NCEE 2010-4013

U. S. DEPARTMENT OF EDUCATION

Compendium of Student, Teacher, and Classroom Measures Used in NCEE Evaluations of Educational Interventions

Volume II: Technical Details, Measure Profiles, and Glossary (Appendices A-G)

April 7, 2010

Lizabeth M. Malone
Charlotte Cabili
Jamila Henderson
Andrea Mraz Esposito
Mathematica Policy Research

Kathleen Coolahan
Gryphon LLC

Juliette Henke
Subuhi Asheer
Meghan O'Toole
Sally Atkins-Burnett
Kimberly Boller
Mathematica Policy Research

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

Compendium of Student, Teacher, and Classroom Measures Used in NCEE Evaluations of Educational Interventions

Volume II: Technical Details, Measure Profiles, and Glossary (Appendices A – G)

April 7, 2010

Lizabeth M. Malone
Charlotte Cabili
Jamila Henderson
Andrea Mraz Esposito
Mathematica Policy Research

Kathleen Coolahan
Gryphon LLC

Juliette Henke
Subuhi Asheer
Meghan O'Toole
Sally Atkins-Burnett
Kimberly Boller
Mathematica Policy Research

NCEE 2010-4013
U.S. DEPARTMENT OF EDUCATION

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences under Contract ED-04-CO-0112/0006.

Disclaimer
The Institute of Education Sciences (IES) at the U.S. Department of Education contracted with Mathematica Policy Research to document and describe outcome measures used in education evaluations. The views expressed in this report are those of the authors and do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

April 2010

Malone, Lizabeth M., Charlotte Cabili, Jamila Henderson, Andrea Mraz Esposito, Kathleen Coolahan, Juliette Henke, Subuhi Asheer, Meghan O'Toole, Sally Atkins-Burnett and Kimberly Boller (2010). *Compendium of Student, Teacher, and Classroom Measures Used in NCEE Evaluations of Educational Interventions. Volume II. Technical Details, Measure Profiles, and Glossary (Appendices A – G)* (NCEE 2010-4013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

This report is available on the IES website at http://ncee.ed.gov.

Alternate Formats
Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at (202) 260-9895 or (202) 205-8113.

DISCLOSURE OF POTENTIAL CONFLICTS OF INTEREST

There are ten authors for this report with whom IES contracted to develop the discussion of issues presented. Drs. Lizabeth M. Malone, Sally Atkins-Burnett, Kimberly Boller, and Ms. Charlotte Cabili, Jamila Henderson, Andrea Mraz Esposito, Juliette Henke, Subuhi Asheer, and Meghan O'Toole are employees of Mathematica Policy Research (Mathematica) and Dr. Kathleen Coolahan, an employee of Gryphon, LLC. The authors and other staff of Gryphon LLC and Mathematica do not have financial interests that could be affected by the content in this report.

CONTENTS

CONTENTS (*continued*)

CONTENTS (*continued*)

CONTENTS (*continued*)

TABLES

EXHIBITS

FOREWORD

The National Center for Education Evaluation and Regional Assistance (NCEE) within the Institute of Education Sciences (IES) is responsible for (1) conducting evaluations of federal education programs and other programs of national significance to determine their impacts, particularly on student achievement; (2) encouraging the use of scientifically valid education research and evaluation throughout the United States; (3) providing technical assistance in research and evaluation methods; and (4) supporting the synthesis and wide dissemination of the results of evaluation, research, and products developed.

In line with its mission, NCEE supports the expert appraisal of methodological and related education evaluation issues and publishes the results through two report series: the *NCEE Technical Methods Reports,* which offer solutions and contribute to the development of specific guidance on state-of-the-art practice in conducting rigorous education research, and the *NCEE Reference Reports*, which advance the practice of rigorous education research by making focused resources available to education researchers and users of education research, to facilitate the design of future studies and help users of completed studies better understand their strengths and limitations.

Subjects selected for *NCEE Reference Reports* are those that examine and review rigorous evaluation studies conducted under NCEE to extract examples of good or promising evaluation practices. The reports present study information to demonstrate the possible range of solutions so far developed. In this way, *NCEE Reference Reports* aim to promote cost-effective study designs by identifying examples of the use of similar and/or reliable methods, measures, or analyses across evaluations. It is important to note that *NCEE Reference Reports* are not meant to resolve common methodological issues in conducting education evaluation. Rather they present information about how current evaluations under NCEE have focused on an issue or on selected measurement and analysis strategies. Compilations are cross-walks that make information buried in study reports more accessible for immediate use by the researcher or the evaluator.

This *NCEE Reference Report* is intended to help researchers select measures for future studies efficiently, assist policymakers in understanding the measures used in existing studies, facilitate comparisons of results across studies, and broaden understanding of these measures within the educational research community.

Selecting outcome measures for use in educational evaluation research is challenging. Researchers face a range of options without having the tools needed to quickly access information about existing and new measures. This report provides detailed, readily accessible, comparative information on the measures that have been used in approximately 40 NCEE evaluation studies to assess student outcomes, instructional practices, teacher pedagogical and/or content knowledge, and classroom environments.

ACKNOWLEDGMENTS

The authors would like to thank other Mathematica staff members including Drs. Aaron Douglas, Sarah Dolfin, and Gail Baxter for their careful review of the measure profiles and of the summary tables of recently developed measures. The Compendium also benefited from Dr. Phil Gleason's thorough review of Volume I and the glossary. In addition, we are indebted to Drs. John Burghardt and Irma Perez-Johnson, who provided input throughout the project and reviewed the Compendium in its draft form. The editing and production staff included Amanda Bernhardt, William Garrett, Cindy George, John Kennedy, Carol Soble, Linda Heath, Cindy McClure, Karen Groesbeck, and Jill Miller.

We would also like to acknowledge the developers of the measures included in this Compendium and are especially grateful to those who took the time to provide updated information and materials on their measures. Their involvement helped ensure inclusion of the most current information on the measures.

Finally, we would like to recognize and refer readers to three other measures compendia that were invaluable resources for several of the measures profiled in the current Compendium and that may add supplementary information to that provided here.

- *Resources for Measuring Service and Outcomes in Head Start Programs Serving Infants and Toddlers* (Kisker et al. 2003) provides descriptions of measures of child outcomes, parent/family background, and program environments as well as a model for measures profiles, for use by infant and toddler practitioners.
- *Early Childhood Measures Profiles Compendium* (Berry et al. 2004) focuses on early childhood measures of children's development and knowledge in a variety of domains.
- *Quality in Early Childhood Care and Education Settings: A Compendium of Measures (Halle and Vick 2007) presents measures to observe early childhood settings (typically for infant, toddler, and preschool settings).*

ABSTRACT

This report contains resources to help researchers and policymakers review measures used in NCEE evaluations of educational interventions. The measures included in the Compendium are applicable to settings for preschool through grade 12. The Compendium discusses criteria and their importance in selecting measures for assessing intervention impacts on student, teacher, and classroom outcomes, and presents profiles or table summaries of these measures. In expectation that the information in this document will be used under diverse circumstances for varied purposes, background information is presented in report format. The materials will be most useful when used in consultation with an assessment expert.

.

APPENDIX A

PROFILE AND SUMMARY TABLE CONTENTS

ROADMAP

Appendix A defines the information in the Compendium's profiles and summary tables of recently developed measures. The Mathematica team developed a standard protocol to summarize information about the measures and to create consistency in presentation across measures. First, the appendix describes the contents of the individual profiles. Second, it reviews the set up of the summary tables for recently developed measures. Subsequent appendixes contain individual profiles as well as a summary table by outcome type—student achievement/development (Appendix B), teacher knowledge (Appendix C), and classroom practices and settings (Appendix D). A glossary of common terms related to assessment, measurement, and psychometrics provides additional information (Appendix E). Finally, Appendix F provides a cross-walk of the full names of the NCEE or REL studies from which the measures were drawn, with the short names used in the Compendium.

PROFILE CONTENTS

A profile contains two components—a one-page summary and a multipage narrative. The summary page provides a quick review of the information considered important in selecting a measure. The narrative details the content, administration, scoring, and psychometrics of the measure. Exhibits A.1 and A.2 (p. A.9) provide a template of the two components that guided profile development.

Profile Summary Page

The first page of the profile provides an at-a-glance summary of the essential characteristics of a measure, including:

EXHIBIT A.1

TEMPLATE FOR COMPENDIUM PROFILE SUMMARY PAGE

Authors:		**Type of Assessment:** **Domain:**
Publisher:		**Grade/Age Range:** **Administration Interval:**
Material, Training, and Scoring Costs:		**Personnel and Training Requirements** **Credentials Required for Use:** **Personnel for Administration:** **Training for Administration:**
Languages:		**Alternate Forms:**
Representativeness of Norming Sample:		**Summary** **Initial Material Cost:** **Time to Administer:** **Ease of Administration and Scoring:** **Reliability:** **Predictive Validity:** **Construct/Concurrent Validity:** **Norming Sample Characteristics:**

- **Authors, Publisher and Material, Training, and Scoring Costs.** The first several sections of the summary page describe how to obtain the measure. The sections list the cost of the initial materials required for use of the measure[1] and, when appropriate, the training costs involved for observer or assessor certification. The sections also present optional training and scoring alternatives.

- **Languages.** The section lists the language(s) in which a measure is available. Some measures may have unofficial translations, that is, translations that usually do not have established psychometric properties and/or do not have publisher approval. The section lists only the languages made available or noted by the authors or publishers. Use of an unofficial translation of a measure may yield scores that are not comparable to the norming sample scores. The narrative section provides more detail on the comparability and equivalence of the English version versus other language versions.

- **Representativeness of Norming Sample.** Knowing whether a measure has norms and, if so, whether the norming sample was nationally representative or representative of the students or teachers under study is an important consideration in selecting a measure to evaluate an educational intervention. If researchers want to know how a given set of students perform compared with students nationally, they must look for a measure with a nationally representative norming sample. The Mathematica team focused on whether the norming sample was nationally representative of students in prekindergarten through grade 12 or a subset of students. The profile refers only to the sample used in the standardization or norming of the measure; that is, it indicates whether the developer clearly states that norms for standard scores, stanines, age equivalents, or similar score transformations are available. The profile notes "No norming sample" and describes the research or pilot samples if, as is commonly the case with classroom observations, some behavior ratings, or direct assessments reporting only the percentage correct, the measure lacks norms.

- **Type of Assessment.** The Mathematica team categorized the measures by respondent (individual providing information) and mode of data collection. Measures typically fell into one of three categories: (1) direct assessment, in which the student or teacher completes a series of items, administered individually or to a group, noting whether administration is adaptive (items tailored to the individual respondent based on the pattern of responses) or not adaptive (all items administered); (2) report or ratings of one's own or another's behavior or knowledge; and (3) observation, in which a trained individual observes the classroom or school and rates or scores the behaviors of interest. Assessment-type categories used during measure reviews included:

  - Group-administered assessment
  - Group-administered adaptive assessment
  - Individual assessment

[1] Some publishers may provide an inspection copy of the materials at no charge for a short period.

- Individually-administered adaptive assessment
- Teacher report (on student outcomes)
- Teacher self-report (on one's own teaching practices or pedagogy)
- Parent report (on student outcomes)
- Parent self-report (on own behavior or outcomes)
- Child self-report
- Classroom observation

- **Domain.** The content covered by the measures fell into several domains applicable to student achievement or development, teacher knowledge, and classroom practices and settings. Table A.1 (p. A.15) lists each domain and a description that the Compendium team developed to categorize the specific measures. Domains and descriptions reflect common terminology in education or subject-specific fields as well as categories developed as part of the Compendium process.

- **Grade/Age Range.** The summary includes the age range for which the measure is appropriate. One measure may assess a broad range of ages and grades or may have several forms for assessing various ages and grades.

- **Administration Interval.** The summary notes the recommended time, if given, between administrations of the measure. Repeat administrations may lead to practice effects (that is, improved performance on an assessment because of familiarity with the items). Some measures, such as classroom observations, may be administered repeatedly without negatively affecting reliability or validity, whereas others (particularly student assessments) are compromised by repeat administrations over a short period. For those that may be administered repeatedly, the administration interval is noted "As frequently as desired." Administration interval may relate to a single form or between administrations of alternate forms, when available.

- **Credentials Required for Use and Personnel and Training for Administration.** Credentials required for use indicate the individual responsible for administrating, scoring, interpreting, and using the results of the measure. Several publishers establish guidelines for who may purchase and interpret the measure. This section of the summary notes whether the publisher/developer requires particular professional or education credentials as a condition of obtaining the measure per the Pearson Assessments publisher qualification levels[2] (see http://pearsonassessments.com/catalog/qualification.htm). For measures not published by Pearson Assessments, the Mathematica team determined whether purchase required specific qualifications and the type of degree or level of experience. For example, the Woodcock-Johnson III requires a "high" credential, which is specified as a master's degree or higher

[2] Pearson Assessments published the majority of the measures that specified qualifications for use. This parent company owns several assessment publishing companies.

with graduate-level training in intelligence/cognitive assessment or neuropsychology.

- In brief, the Compendium credential levels included Level 1 for requiring training or supervised experience with measurement, Level 2 for holding a bachelor's degree with coursework in measurement and domain, and Level 3 for requiring licensure or state certification or a doctoral degree. The Mathematica team added a Level 2+ to connect to other qualification rankings that fell between bachelor's and doctoral degrees (for example, a master's degree). All levels should be considered approximations; publishers may grant special permission at any level. Otherwise, the summary notes when no special qualifications are required or when qualifications information is not specified.
- The personnel who administer the measure may differ from the researcher who purchases the measure. Under the researcher's supervision, the personnel hired and the training provided by the developer vary with the requirements of a given measure. Personnel ratings include whether the measure requires administration by a researcher with specialized training, a highly trained staff member, or a clerical staff member. The Mathematica team also included an estimate of how much time such a person would need to learn, conduct, and score the measure. Some authors and publishers recommend that an experienced assessor observe or review administrations conducted by trainees. The profiles include such information along with the availability of or recommendation for group training on the use of the measure and the training cost.

- **Alternate Forms.** Some publishers provide two or more versions of the same measure so that the same skills or behaviors may be assessed several times with reduced concern that scores may change as a result of "learning the test" from repeated administration of the same items. The profile indicates the existence of alternate forms and any evidence that they have been tested for equivalence (and considered interchangeable). The profile also notes how many alternate forms are available and the recommended administration interval. Alternate parallel forms provide the researcher with options to reassess a construct with the same measure over shorter intervals with less concern for practice effects (that is, improved performance associated with familiarity with the items over repeated test administrations).

- **Summary.** The summary section of the first page of the profile presents key features to help the reader make comparisons across measures. However, the categorical ratings do not reflect a recommendation of any particular measure. An individual must determine which features are most important for the study purposes, sample, and context. Features included in the summary section are:

  - **Initial material cost.** Material cost is 1 (under $100), 2 ($100 to $200), 3 ($200 to 500), or 4 (> $500) based on the cost of an assessment kit or, if no kit is required, the cost of forms and manuals and the cost of any easily

identifiable training or scoring costs reported by the authors or publishers. For some measures the Mathematica team did not present a fixed price for materials because costs may depend on particular study parameters (for example, number of data collection waves or number of students/teachers to be assessed). In these instances, the Mathematica team rated cost as "To be determined based on negotiations with the publisher," or TBD.

- **Time to administer.** Administration time is the number of minutes or hours that the measure requires for completion; if administration times vary, the reader is directed to the Description in the narrative for details.
- **Ease of administration and scoring.** The administration and scoring process is 1 (not described), 2 (self-administered or administered and scored by someone with basic clerical skills), 3 (administered and scored by a highly trained individual), or 4 (administered or scored by a clinician or specialist). Some measures may have requirements that vary between administration and scoring; in such cases, the higher rating applies.
- **Reliability.** Reliability is 1 (none described), 2 (all or mostly under 0.70), or 3 (all at or above 0.70). For direct assessments of knowledge and reports of behavior, the reliability category relied on internal consistency estimates that the items in the measure capture the same construct. For observation tools, the reliability category reflects the inter-rater reliability estimates of consistency that provide evidence that the tool captures the same information across observers. We chose the threshold of 0.70 following the prevalent rule of thumb in the field used by researchers and assessment developers.[3] Reliability, a prerequisite for validity, indicates the consistency and stability of a measure. Other things being equal, the higher the reliability, the better the measure is. If the reliability category (2 versus 3) varied across total scores and composite scores, the category in the summary section focuses on total scores, with a reference note indicating variability.
- **Predictive validity.** "Available" or "Not Available" indicates whether information exists about the relationship between the measure and another measure or criterion administered later. See the profile's narrative section on Validity Evidence for more details on information captured.
- **Construct/concurrent validity.** "Available" or "Not Available" indicates whether information exists on the measure's convergence with other measures of the same construct or the measure's concurrence with other measures of the same or a similar construct or related criteria (at the same time). See the

[3] For observation measures, inter-rater reliability was considered as a category 2 if 1) percentage agreement (exact or within a certain point range was 85 percent, 2) kappa coefficient was 0.60 or (3) the intraclass correlation coefficient was 0.80. For the reading fluency measures, developers typically do not conduct estimates of internal consistency reliability. Some developers note such an estimate is not appropriate for timed fluency measures, while other researchers indicate the potential to do so. The current compendium selected internal consistency reliability for determining the summary reliability rating, and thus, the reading fluency measures received a rating based on the presence (or absence) of that information.

profile's narrative section on Validity Evidence for more details on information captured.

- **Norming sample characteristics.** The designations 1 (none described), 2 (older than 10 years or not nationally representative), and 3 (normed within past 10 years and nationally representative) indicate norming sample characteristics. Many publishers/developers renorm or standardize their measure every 10 years to ensure that the measure is representative and current. Norming samples that are weighted to be nationally representative receive a 3 if data were collected within the past 10 years, with a note added to the rating to identify this approach.

Profile Narrative

Following the one-page summary with the information described above, a multipage narrative details the measure's administration, scoring, interpretation, and technical properties. The narrative provides the same type of information as that captured in the profile summary page but provides more detail and includes a description of content, uses of information, training, and adaptations. Exhibit A.2 (p. A.9) presents the format and content of the narrative.

- **Description.** This section provides an overview of what the measure was designed to assess and with whom and how the information is collected. In particular, it notes type of assessment, content measured, appropriate grade range, type of stimulus, number of items, average administration time, and further detail on content such as availability of subtests. For adaptive measures, the section describes basal and ceiling rules for administration to determine starting and stopping points.

- **Other Languages.** If a measure is available in a language other than English, this section identifies the language, measure name, and norming sample. It also presents, if available, information from publishers/developers on the comparability of scores from different language versions. In keeping with psychometric terminology, this section notes whether the versions are parallel (statistically equal), equivalent (not statistically similar but compensated for in score conversion), or comparable (no demonstrated statistical similarity).

- **Uses of Information.** Publishers/developers design a measure for a given purpose and usually describe the intended and appropriate uses of the measure as a clinical or research tool. To support users in determining whether a measure's intended use is aligned with their goals, this section summarizes how publishers/developers characterize their measures. Measures may be used to assess status or growth. Some are designed to screen students while some capture an in-depth assessment of a particular domain. Some provide feedback on classroom procedures as part of a quality improvement process.

- **Methods of Scoring.** Student achievement or teacher knowledge measures may be scored by using a point system and summing correct responses; reports of behavior or classroom observation measures may use a broader range of response categories,

such as whether a particular behavior occurs with varying frequency or intensity. This section summarizes the response options, along with procedures for scoring a measure and the types of scores it is possible to compute.

- **Interpretability.** Many developers provide information about how to interpret scores or ranges of scores derived from their measures. This section summarizes information available to assist in interpreting a measure and specifies the education and experience level required of the person who interprets the results.

EXHIBIT A.2

TEMPLATE FOR COMPENDIUM PROFILE NARRATIVE

NARRATIVE

Description:

Other Languages:

Uses of Information:

Methods of Scoring:

Interpretability:

Reliability:
(1) Internal consistency reliability:
(2) Test-retest reliability:
(3) Alternate form reliability:
(4) Inter-rater reliability:

Validity Evidence:
Construct/Concurrent validity:
Predictive validity:

Bias Analysis:

Training Support:

Adaptations/Special Instructions for Individuals with Disabilities:

Alternate Forms:

Previous Version:

NCEE or REL Study Use:

References:

Reliability. A reliable assessment is dependable and stable; that is, the results are similar when administered to the same individual on multiple occasions over a period of time. Overall, the lower the reliability of a measure, the greater is the error associated with the measurement. The error may originate from the items, the timing of the assessment, how the assessment is administered, or other sources. Indicators of reliability examine internal consistency reliability, test-retest reliability, alternative form reliability, and inter-rater reliability. Correlations and coefficients of reliability estimates presented in the profiles are assumed to be significant ($p<0.05$) unless otherwise noted; if a measure's manual did not provide *p*-values, the Mathematica team assumed that the developer reported only significant values. For correlations, some researchers provide a common standard for weak correlations as below 0.30. The information presented about correlations is uncorrected when available, but the Mathematica team notes the presentation of corrected correlations (and the reason for the correction; for example, measurement error or restricted range); such corrections may increase the magnitude of the correlation and not reflect original data. The types of reliability summarized in the Compendium profiles include:

- **Internal consistency reliability.** Coefficients (split-half reliability, Cronbach's alpha, Kuder Richardson-20, or Item Response Theory [IRT] reliability estimates) indicate the extent to which the items in the measure "hang together"; all of the items seem to provide information about and measure the same construct.
- **Test-retest reliability.** This type of reliability indicates the extent to which a measure yields the same results when administered to the same test takers at different times. The interval between administrations, typically under one month, is noted when available.
- **Alternate form reliability.** This type of reliability indicates the extent to which the measure yields the same results when a different form of the measure is administered (for those with several forms), usually by using a split-half reliability coefficient or bivariate correlation.
- **Inter-rater reliability.** This type of reliability measures the extent to which two observers or assessors would interpret and record a given set of information in the same way. Reliability coefficients include a correlation coefficient, Kappa coefficient, intraclass correlation coefficient (ICC), or generalizability coefficient. Developers may also report the percentage agreement (exact match or within a certain point range on a scale). Measures that involve self-report or ratings of behavior or beliefs do not necessarily have a right/wrong answer such that inter-rater reliability information about them is generally not applicable. Intra-rater reliability is another type of reliability that applies to a single observer of the same scenario at different points to assess coder drift or realignment to coding standards.

Validity Evidence. Indicators of validity help determine whether a measure assesses what it is supposed to assess for its intended purpose. For example, if a measure purports to provide an estimate of a student's vocabulary, how the student performs on the

measure should be similar to how he or she performs on another established vocabulary assessment. The Mathematica team assumed that, unless clearly noted by the authors/publishers, correlations between two measures meant to provide validity evidence were significant ($p<0.05$). Similarly, unless otherwise noted, the correlations—when available—represent uncorrected estimates. The presentation of corrected correlations (and the reason for the corrections; for example, restricted range or measurement error) is noted because such corrections may increase the correlation's magnitude and not reflect original data. Validity evidence includes documentation of the alignment of the measure with the overarching construct that the measure is intended to assess. As discussed in Volume I, Chapter II, validity may be categorized in different ways. The Compendium presents three types of validity evidence:

- **Content validity.** This type of validity describes the evidence of how well the content of the measure represents the relevant aspects of the construct. Developers often use expert judgment to determine that a measure includes what it is supposed to include and that the content is relevant. As part of the content validity description, the Mathematica team describes sources used to develop the measure and any research or literature noted as a foundation for developing the measure. For measures without a norming sample (as noted in the profile summary page), the content validity section describes the research sample and any pilot work.
- **Construct/concurrent validity.** This type of validity provides information on analyses conducted to examine the structure of the measure's items or scales and comparisons of a measure with other measures to support the construct and/or concurrent criterion-related validity.

  Analyses that establish scales or subtests as distinct dimensions of a construct include IRT analytic approaches and exploratory or confirmatory factor analysis. The analyses also provide construct validity information about test functioning and relevant item statistics. In particular, item difficulty statistics note the probability of answering an item correctly and provide information on the range of responses covered. Item discrimination statistics communicate how well a change in ability predicts a likely correct response to a particular item.

  Comparisons of a measure to other measures or criteria capture the meaningfulness of a measure in practical use. Three primary types of validity evidence compare a measure's results to other information collected at the same time: (1) convergent validity compares measures of the identical construct; (2) criterion-related validity compares measures of a similar construct or indicator (such as graduation) that are expected to be related; and (3) divergent or discriminant validity compares measures of different constructs that are not expected to be related and thus would have a weaker or no correlation with each other.[4]

[4] A comparison of scores from measures of different constructs would examine correlation coefficients for low or zero values (or potentially negative values). In cases where the comparison is between subtest scores within a

The Mathematica team also includes developers' work on discriminant analysis to provide evidence that a measure can distinguish groups as expected. For example, for a measure that is a screener, it should distinguish students with or without disabilities; for a measure of development, it should demonstrate an expected difference between means at different ages or grades. Analyses may show differences in mean scores, sensitivity to the correct identification of students, and specificity or the rate of "true negatives."

- **Predictive validity.** This type of validity indicates the extent to which the measure's results are related to later functioning. Consequently, new measures do not provide this information. If the manual contains information on a previous version of the measure for predictive validity, that information provides a basis for reporting. Any consideration of predictive validity evidence from previous versions should take account of whether the most recent version is revised or only renormed and then look for any evidence of the relationship between performance on the previous and current versions.

- **Bias Analysis.** Developers may undertake several activities and/or analyses to determine whether items function differently for particular groups. Often developers convene expert panels to review items for bias (for example, cultural, gender, or socioeconomic bias). Appropriate analyses include an examination of items for differential item functioning (DIF) for particular subgroups; however, not all developers have necessarily undertaken DIF. Therefore, the profiles include demographic subgroup comparisons for mean scores and internal consistency coefficients. The glossary (Appendix E) provides further information about bias analysis and DIF.

- **Training Support.** To help researchers identify staffing and training needs and resources, this section summarizes the authors'/publishers' recommendations about their measures. The profiles describe the available training materials, products, or sessions. Some authors and publishers include a great deal of information on preparation for administering a measure; others provide little information.

- **Adaptations/Special Instructions for Individuals with Disabilities.** Some measures are designed specifically to assess the abilities or performance of individuals with disabilities, but most are not. The profiles describe adaptations or instructions the authors or publishers included for working with people with disabilities. For example, for student outcome measures, the assessor may administer the measure differently to students with disabilities based on standardized procedures documented by the developer, or alternative versions in Braille or large print may be available. If students with disabilities were not included in the norming sample or procedures were not standardized during the norming process, the scores obtained for students with disabilities may not be comparable to norms.

(continued)

single measure, the observed scores demonstrate shared method variance (from the same measure) such that it is reasonable to expect the correlation to seem higher than expected for divergent or discriminant validity evidence.

- **Alternate Forms.** This section notes the availability of alternate but comparable forms, along with how many exist and the recommended length of time required between administrations of the alternate forms.

- **Previous Version.** Although the profile reviews the most recent version of a measure, some studies may have used a previous version of a measure based on availability or comparability to other studies or previous waves. For example, the Study of Classroom Literacy Interventions and Outcomes (CLIO) in Even Start collected data from 2003 through 2006 using the Peabody Picture Vocabulary Test, Third Edition (PPVT-III). In 2007, a fourth edition of the PPVT was released and is the current measure available on the market with the latest norms (if available). Therefore, the Mathematica team notes differences between the latest version profiled in the Compendium and the previous version. This information may help researchers and policymakers determine the comparability of outcome measures across studies.

- **NCEE or REL Study Use.** The inclusion criteria for the Compendium focused on outcomes of educational intervention evaluations conducted by an Institute for Education Sciences (IES), National Center for Education Evaluation and Regional Assistance (NCEE), or Regional Educational Laboratory Program (REL); the evaluations used a rigorous experimental or quasi-experimental design. Some studies have used the measures either intact or modified. This section names the studies using the measure (or a previous version of it), with a hyperlink to the study NCEE or REL web page or most recent report.

- **References.** Profiles include full citations for manuals and other sources of information as well as citations for any other materials the authors/publishers make available about the measure, such as training videotapes and computer scoring programs and materials for Spanish (or other language) versions of the assessment.

SUMMARY TABLE OF RECENTLY DEVELOPED MEASURES

Several NCEE or REL studies use new measures, often developed by a study, to focus on particular student, teacher, or classroom outcomes when available measures do not meet study needs. Because these measures are still in development or were recently released, limited information is available on their psychometric properties (a requirement from the outset for profiles). Thus, we have developed summary tables for these more recently developed measures that are not as detailed as the profiles but nonetheless provide some information about the measures. The Compendium includes these measures so that future studies may build on current studies.

The Mathematica team compiled the summary tables from NCEE or REL study reports and with some communication with study directors. Given that many measures are part of ongoing data collection efforts funded through a federal contract, the most common information source was Office of Management and Budget (OMB) clearance packages required for conduct of a study. OMB packages include a justification for and description of each measure proposed for a federally funded study. The summary table entry for each newly developed measure includes six columns detailing:

1. **Measure.** The measure's name is the name included in the study materials and may be just a general reference to a questionnaire or classroom observation. We distinguish such names by adding short phrases on key content included in the questionnaire or observation.

2. **NCEE or REL Study Use.** This section names the study using the measure, hyperlinked to the study's web site or report.

3. **Description.** An overview of the measure includes the domains (Table A.1) covered by the content, type of assessment (matching the options used in full Compendium profiles; see above), specific constructs assessed, and other details if available (for example, number of items, types of scores, and administration time).

4. **Grade/Age Range.** This specifies the age(s) for which the measure is appropriate and is generally expressed as the grade or age of the study sample.

5. **Reliability.** Correlations and statistics investigating the measure's reliability take the form of internal consistency reliability, test-retest reliability, alternate form reliability, and inter-rater reliability, with specific types and values noted as available. Some measures (especially student and teacher questionnaires) include items from previously developed scales and large-scale surveys that have been modified. In these cases, study researchers (especially in the early design stages) may note reliability information for the original measures, as reflected in the summary table.

6. **Validity.** Evidence must indicate that the measure assesses the construct it purports to assess, particularly as related to content, construct, concurrent, and predictive validity. Most commonly, the entry is either "No Information Available" or the content validity evidence is from the measure's development phase. If study researchers drew items from other scales and large-scale surveys, such sources are noted.

SUMMARY

Together, the profile and summary tables provide a detailed review of measures for assessing the impact of educational intervention evaluations on student achievement/ development, teacher knowledge, and classroom practices and settings. The content described above represents key areas for consideration when reviewing a measure for selection for a given study. The information collection method relied on a consistent set of sources to locate information (see Volume I, Chapter III). The application of the content as described in the current appendix ensured a standard format for summarizing the information across measures.

TABLE A.1

STUDENT, TEACHER, AND CLASSROOM DOMAINS OF MEASURES INCLUDED IN THE COMPENDIUM

Domain	Description
Student Achievement/Development	
Reading	Early literacy skills such as phonological awareness, letter recognition and naming, and print concepts as well as reading vocabulary, decoding, phonics, reading fluency, and various comprehension skills.
Language arts/language proficiency	Assessment of expressive and receptive language skills (oral and written). Areas such as writing, oral skills, editing skills, grammar, syntax, vocabulary, and morphology may be included in these assessments along with English language proficiency or Spanish language proficiency.
Mathematics	Skills and topics related to counting, calculation, problem solving, geometry, and algebra.
Science	Fields of earth, space, physical, and life sciences.
Social studies	Topics across a range of disciplines such as history, culture, geography, and economics.
Approaches to learning/motivation	Includes executive functioning, attention, cognitive flexibility, curiosity, and engagement in learning.
Social-emotional	Includes social skills and problem behaviors.
General knowledge	Includes understanding of spatial relations and knowledge of colors.
Others specified as needed	Areas such as motor development and substance use.
Teacher Knowledge	
Subject knowledge	Knowledge about content, topics, constructs, and procedures and a specific subject, noting the particular content subject area.
Pedagogical knowledge	Knowledge about how to teach in general.
Pedagogical content knowledge	Knowledge about how to teach the content of a particular subject or domain of learning, noting the particular content subject area.
Classroom Practices and Setting	
Classroom quality	Aspects and quality level of physical, social, and temporal environments to include effective classroom management, use of routines, time use, interactions, materials, and space. The profile notes in parentheses if a particular measure assesses teacher-student interactions or the classroom environment. Teacher-student interactions include areas such as positive support, warmth, negative interactions, and punitiveness. Environment rates the materials and space available for learning.
Instructional practices	The practices, activities, and strategies employed in the teaching of students and including both teacher-initiated instructional practices and feedback, noting whether the measure is comprehensive or subject-specific. Comprehensive measures cover all subject areas in giving a rating; subject-specific refers to instructional practices that address a particular domain of learning.
School climate	Sense of safety, positive regard for members of the community (class or school), and sense of community. Example items include "I (student) feel safe at this school" and "The students at this school work well together."
School engagement	Measures of student motivation, engagement, or involvement at an aggregate level (students in this classroom, students in this school).
Motivation for teaching	Includes teacher enthusiasm, self-efficacy, and confidence.

APPENDIX B

STUDENT ACHIEVEMENT AND DEVELOPMENT MEASURE PROFILES AND TABLE OF RECENTLY DEVELOPED MEASURES

6+1 TRAIT WRITING SCORING GUIDE (RUBRICS), 2004

Authors: Northwest Regional Educational Laboratory (NWREL)		**Type of Assessment:** Individual or group-administered assessment **Domain:** Language arts/language proficiency (writing)
Publisher: Northwest Regional Educational Laboratory (NWREL) 800-547-6339 http://www.nwrel.org		**Grade/Age Range:** Grade 3 through 12 **Administration Interval:** As frequently as desired
Material, Training, and Scoring Costs: The Scoring Guide and training exercises are available on NWREL's web site for free. Optional "Trait-Tote" kits for grades 3 through 5 and grades 6 through 8 (reference and guide books, DVD, sample papers, parent handbook, and instructional aids): $99.00 each		**Personnel and Training Requirements** **Credentials Required for Use:** Level 1 (training or supervised experience with measurement) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (> 2 hours)
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (under $100) **Time to Administer:** Not specified **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The 6+1 Trait Writing Scoring Guide is an assessment tool for use with the 6+1 Trait Writing for Assessment and Instruction model (referred to in abbreviated form as the 6+1 Trait writing model). Developed by the Northwest Regional Education Laboratory (NWREL), the model is an analytic instructional approach that employs formative assessment to inform classroom writing instruction. The scoring guide, often referred to as the 6+1 Trait rubrics, is generally used with students in grade 3 through 12.[1]

The rubrics may be used to rate student writing with respect to six core traits[2] associated with writing achievement and quality: ideas (content or message), organization, voice (expression of the writer's feelings and convictions), word choice, sentence fluency, and conventions (grammar and mechanics). Presentation (relating to form and layout) is an optional seventh trait. Each trait has a separate scoring rubric that includes a continuum of quality ratings, ranging from 1 (low quality) to 5 (high quality). The continuum lists main descriptors characterizing the trait-based criteria associated with points 1, 3, and 5 on the continuum. For example, on the ideas trait, the main descriptor associated with a score of 5 reads: "This paper is clear and focused. It holds the reader's attention. Relevant anecdotes and details enrich the central theme." In contrast, the main descriptor for a score of 1 is: "As yet, the paper has no clear sense of purpose or central theme. To extract meaning from the text, the reader must make inferences based on sketchy or missing details. The writing reflects more than one of these problems [lists problems]." Beneath each main descriptor, five criteria further define the trait qualities associated with a particular score. For all traits, a score of 3 represents a developmental midpoint at which strengths and weaknesses are equally balanced with respect to the trait. Based on the descriptors and criteria for points 1, 3, and 5, assessors are expected to infer the qualities associated with ratings of 2 and 4.

Other Languages: None.

Uses of Information: The information yielded by the 6+1 Trait rubrics may serve several purposes. First, it provides feedback to teachers and students about the quality of student writing as assessed on each trait, indicating strengths and weaknesses. Second, the developers state the rubrics may be used as research and evaluation tools to assess student achievement in writing across time and to assess and compare the effectiveness of writing instructional methods. Third, as part of an instructional model, the information may be used by teachers for planning and delivering writing instruction.

Methods of Scoring: To use the scoring rubrics, the assessor assigns a score from 1 to 5 for each trait. Trait scores are not averaged or combined into a total score. Teachers who wish to convert trait scores to grades may use the 6+1 Trait Rubric to Grade Converter (sold by the publisher), a slide chart with suggested grades for each trait based on trait scores. No information is provided on how the grade conversion was determined.

Interpretability: The 6+1 Trait rubrics are designed to quantify strengths and weaknesses in each trait. Scores for each trait are interpreted separately. Higher ratings indicate more highly developed writing.

Reliability:
(1) Internal consistency reliability: No information available.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: Kozlow and Bellamy (2004) reported Cohen's Kappa coefficients ranging from 0.96 to 0.99 as estimates of inter-rater reliability and a rate of scoring discrepancies ranging from 0.50 to 9.0 percent. Two assessors used the 6+1 Trait rubrics with a sample of 1,592 students in 72 classrooms (grade 3 through 6).

Validity Evidence:
The 6+1 Trait Writing model and scoring rubrics are based on decades of research on the writing process and instruction, the use of formative assessment to inform instruction and enhance student achievement, and cooperative learning. Kozlow and Bellamy (2004) collected validity evidence on the rubrics from a sample of 1,592 students in grade 3 through 6 in 72 classrooms in a single school district. The sample consisted of almost all White students who were native English speakers (fewer than 1 percent of sample students were racial/ethnic minorities, and fewer than 1 percent were English Language Learners). Ten percent of sample students were eligible for free or reduced-price lunch, and 11 percent were students with disabilities.

Construct/Concurrent validity: For each grade level, Kozlow and Bellamy (2004) calculated nonparametric correlation coefficients (Kendall's tau) to investigate relationships among scores on the first six 6+1 Trait rubric scores. They reported that the nonparametric coefficients, used because the rubric scales are not interval scales, were about 0.1 lower than the product-moment coefficients at all grade levels. Inter-correlations among the six trait scores ranged from 0.53 to 0.69 for scores on grade 3 writing samples and from 0.33 to 0.69 for scores on grade 4 through 6 writing samples. At all grade levels, the highest coefficients were between word choice and sentence fluency and between sentence fluency and conventions. The lowest correlations were between conventions and ideas and between conventions and voice, indicating that the traits represent separate features of writing.

Kozlow and Bellamy (2004) also calculated Kendall's tau coefficients to estimate relationships between the 6+1 Trait scores and scores on a holistic scoring rubric that was used by separate raters to assign an overall score to each student writing sample. Correlation coefficients between trait scores and holistic scores ranged from 0.54 to 0.64 for scores on grade 3 writing samples and from 0.40 to 0.57 for scores on grade 4 through 6 writing samples.

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: NWREL conducts workshops at local and NWREL sites to train individuals in implementation of the 6+1 Trait Writing for Assessment and Instruction model. Workshops offer one or two days of training in (1) evaluating student writing with use of the scoring guide to rate performance according to the traits and (2) implementing trait-based classroom instruction. NWREL also conducts training of trainers programs.

Other types of training support include "Trait-Tote" kits that are available for purchase from NWREL. The kits contain training books and DVDs as well as scoring practice materials. In addition, the trait scoring guide, scoring examples, and practice exercises are available at no charge on NWREL's web site.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: None.

Previous Version: None.

NCEE or REL Study Use:[3] An Investigation of the Impact of the 6 + 1 Trait® Writing Model on Student Achievement

[1] The NWREL has also developed a Beginning Writer's Continuum (BWC) that is designed to document students' acquisition of early writing skills in early elementary school.

[2] In addition to the six core traits, some studies have employed a holistic scoring rubric (Kozlow and Bellamy 2004; Northwest Regional Educational Laboratory 2007).

[3] See Table F.1 for web address.

References:

Culham, Ruth. *6+1 Traits of Writing: The Complete Guide for the Primary Grades*. New York: Scholastic, 2005.

Culham, Ruth. *6+1 Traits of Writing: The Complete Guide (Grades 3 and Up)*. New York: Scholastic, 2003.

Culham, Ruth. *Traits of Writing: A Professional Development Video Series on DVD: DVD Version*. New York: Scholastic, 2006.

Kozlow, Michael, and Peter Bellamy. "Experimental Study on the Impact of the 6+1 Trait Writing Model on Student Achievement in Writing." Portland, OR: Northwest Regional Educational Laboratory, 2004.

Northwest Regional Educational Laboratory. "6+1 Trait® Writing Scoring Guide." Available at [http://www.nwrel.org/assessment/scoring.php?d=9]. 2005.

Northwest Regional Educational Laboratory. "6+1 Trait® Writing Scoring Guide--5 Point Condensed." Available at [http://www.nwrel.org/assessment/scoring.php?d=9]. 2007.

Northwest Regional Educational Laboratory. "OMB Supporting Statement Part B: An Investigation of the Impact of a Traits-Based Writing Model on Student Achievement." No. (03299)1850-0835-v.1. Available at [http://edicsweb.ed.gov/browse/browsecoll.cfm?pkg_serial_num=3299]. March 2007.

AIMSWEB ORAL READING FLUENCY, 2002

Authors: Mark R. Shinn and Michelle M. Shinn		**Type of Assessment:** Individual assessment **Domain:** Reading (reading fluency)
Publisher: Pearson 866-313-6194 http://www.aimsweb.com		**Grade/Age Range:** Grades 1 through 8 **Administration Interval:** Standard Benchmark Reading Assessment Passages: 3 times per school year (fall, winter, and spring); Standard Progress Monitoring Reading Assessment Passages: Frequently throughout the school year
Material, Training, and Scoring Costs: The downloadable Oral Reading Fluency components include 24 Benchmark Passages and 250 Progress Monitor Passages across grades 1 through 8 and primer level:[1] $138 (with individual license for 30 or fewer students); $398 (with school license) Training materials include Training Workbook, Technical Manual, 8 video examples, and PowerPoint overview of the measure: Free (publisher's web site) AIMSweb offers several general trainings:[2] Onsite 2-day workshop for 30 participants for $3,700 (with training materials); open 2-day workshop for assessor and 2 guests for $349 (with training materials); and 3- to 5-hour online training for $299 per person		**Personnel and Training Requirements** **Credentials Required for Use:** Level 1 (training or supervised experience with measurement) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours) Assessors must study the training materials and conduct inter-rater reliability checks (see Training Support below).
Languages: English, Spanish		**Alternate Forms:** Three Benchmark Standard Reading Assessment Passages and 20 to 30 Standard Progress Monitor Reading Assessment Passages
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 3[3] ($200 to $500) **Time to Administer:** About 5 minutes **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 1[4] (none described) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Oral Reading Fluency test, a reading curriculum-based measurement (R-CBM) tool, is an individually administered timed assessment that measures general reading achievement for students in grades 1 through 8. Students are given one passage (at their grade level) to read aloud during a one-minute period to determine the number of words read correctly (WRC). Three Standard Benchmark Reading Assessment Passages (RAP) are provided for each grade (24 total) and administered throughout the school year (1 passage administered in fall, 1 in winter, and 1 in spring); 20 (grade 1) to 30 (grade 2 through 8) Standard Progress Monitor RAPs may be administered throughout the year as needed. The assessor must also complete a Qualitative Features checklist following the administration of 3 Standard RAPs (Benchmark and/or Progress Monitor) to assess the quality of the student's reading skills. Including the instructions, testing, and scoring, administration time is about five minutes.

Other Languages: The Spanish version of the Oral Reading Fluency is available for purchase (Benchmark RAPs only). The Training Workbook for the Spanish version will be available at a later date, although assessors may refer to the English version for administration instructions. The Spanish version of the Oral Reading Fluency refers to the same Technical Manual as the English version, but it is not clear if the pilot study included Spanish-speaking students.

Uses of Information: The authors state that the Oral Reading Fluency measure may be used to measure student reading growth and development as well as reading comprehension. The authors note that the Oral Reading Fluency measure is on an approved list for screening as part of the federal Reading First legislation (Shinn and Shinn 2002).

Methods of Scoring: Administration begins when the first word is spoken and ends after one minute (a stopwatch should be used). The assessor scores the measure by marking a slash (on the assessor copy of the reading passage) through words pronounced incorrectly and summing the number of WRC and incorrectly within one minute. Scores are reported as words read correctly to errors (WRC/Errors). For example, for a student with 142 WRC and 3 Errors, his/her score is reported as 142/3. The appendix to the Training Workbook provides detailed information on scoring rules, such as how to score omitted words and self-corrections.

Interpretability: To calculate the percent of WRC, the assessor must first note the total number of words in the passage (listed on the assessor copy of the passage). The assessor then needs to complete the Qualitative Features checklist, which is in the appendix of the Training Workbook, after administering three Standard RAPs (Benchmark and/or Progress Monitor). The checklist indicates whether a student displays a series of reading skills, such as reads very accurately (greater than 95 percent WRC) and demonstrates fluid reading skills.

Reliability:
(1) Internal consistency reliability: No information available.
(2) Test-retest reliability: Four studies conducted test-retest reliability across 2-, 5-, and 10-week periods. In the first study, the correlations for scores between administrations had a mean of 0.90 across grades 3 through 6. In the second study, the correlations between scores of the administrations had a median of 0.90 for a sample of grade 5 students (sample size unclear). In the third study, the correlation between scores of the two administrations was 0.97 for a sample

of 30 grade 5 students. In the last study, the correlation between scores of two administrations was 0.92 for 566 students in grades 1 through 6. The types of passages were not specified.
(3) Alternate form reliability: Alternate form reliability was conducted within the same testing period or within a one-week period across the same four studies. In the first study, the correlations between scores of the alternate forms ranged from 0.84 to 0.96, with a mean of 0.91 across grades 3 through 6. In the second study, the correlations between scores of the alternate forms ranged from 0.89 to 0.94 for a sample of grade 5 students (sample size unclear). In the third study, the correlation between scores of the alternate forms was 0.94 for 110 grade 4 students. In the last study, the correlation between scores of the alternate forms was 0.89 for 566 students in grades 1 through 6. The types of passages were not specified. The average Standard Benchmark and Progress Monitor RAP grade correlations ranged from 0.80 (grade 7) to 0.90 (grade 8), using 23 passages for grade 1 and 33 passages for each of grades 2 through 8.
(4) Inter-rater reliability: Inter-rater reliability was 0.99, using 566 students in grades 1 through 6 (the number of raters and passages were not specified).

Validity Evidence:
Nine teachers and seven paraprofessionals from medium-sized suburban and rural education districts in the Midwest were trained to create the reading assessment passages. Guidelines specified the length of passages for each grade and the number of syllables and sentences per 100 words for each grade based on the Fry readability formula. Three methods—alternate form reliability, standard error of measurement (SEM) comparisons, and Lexile-graded standards—determined the passages for elimination based on approximately 20 students per grade from suburban and rural Midwestern school districts in February and March 2001. Ten of 33 passages were dropped for grade 1 and 17 of 50 passages for grades 2 through 8 based on alternate form reliability estimated at or below 0.70 or passage mean WRC scores more than +1 SEM outside the mean for the grade. Passages that did not receive a Lexile score consistent with expectations for the grade level were also dropped.

Construct/Concurrent validity: The Oral Reading Fluency passages were correlated with passages from the Comprehensive Tests of Basic Skills (CTBS) (e.g., MacMillan Series r), achievement assessments (e.g., California Achievement Test), and other reading measures (e.g., Woodcock Reading Mastery Test). Most correlations were between 0.50 and 0.91 across studies, and four were below 0.50 (correlations were 0.26, 0.39, 0.40, and 0.41 between the Oral Reading Fluency and the Holt, Rinehart, & Winston Basal Readers (grade 4); Harcourt-Brace-Jovanovich (HBJ) Basal Reader (grade 5); Holt, Rinehart, & Winston Basal Readers (grade 6); and the Kaufman Test of Educational Achievement-Brief (KTEA-B) (grade 4), respectively). Sample sizes ranged from 21 to 479 students across grades 1 through 8 for the 20 studies used.

The developers also used Lexile-graded standards, which estimate reading passage difficulty, and readability formulas, which calculate readability. Correlations between the Standard Benchmark RAPs and Lexile-graded standards and readability formulas ranged from 0.78 to 0.99, with a median of 0.95. Correlations between the Standard Progress Monitor RAPs and Lexile-graded standards and readability formulas ranged from 0.78 to 0.98, with a median of 0.90.

Mean words read correctly using the Benchmark and Progress Monitor RAPs indicated a direct relationship between age and performance on the Oral Reading Fluency test, with means

increasing with age (35.7 in grade 1 to 154.2 in grade 7) based on 254 RAPs (23 in grade 1 and 33 in grades 2 through 8). The grade 8 mean WRC was 147.3.[5] Based on the three Standard Benchmark RAPs, mean WRC ranged from 36.2 (grade 1) to 154.1 (grade 7). The grade 8 mean WRC was 147.2.[5]

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: The AIMSweb web site provides several training aids for the Oral Reading Fluency assessment, including the Training Workbook, Technical Manual, research studies, eight sample video administrations (video links require QuickTime software) and assessor copy passages (including an answer key and summary of scoring rules and examples), as well as a PowerPoint overview of the Oral Reading Fluency. Instructions for determining inter-rater reliability are in the Training Workbook and PowerPoint presentation (a formula to determine inter-rater reliability is provided, but no benchmark is specified). To ensure assessor consistency in administering and scoring the assessment, the authors recommend that experienced assessors observe trainees administering the Oral Reading Fluency. In addition, the assessor should complete the Accuracy of Implementation Rating Scale (AIRS) (in the appendix of the Training Workbook), which is a checklist to determine accuracy in trainee administration. AIMSweb offers several general trainings, including on-site two-day workshops for up to 30 participants ($3,700, including training materials), open two-day training workshops for the assessor and two guests ($349, including training materials), and a three- to five-hour online training ($299 per person). The trainings are not required for administration of the Oral Reading Fluency. All of the Oral Reading Fluency training aids (except the Technical Manual) may be found at http://www.aimsweb.com/support-training/training/training-materials/. The Technical Manual may be found on the main AIMSweb Oral Reading Fluency web page under the Description tab or at http://www.aimsweb.com/uploads/pdfs/passagestechnicalmanual.pdf.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: The 3 Benchmark Standard RAPs within each grade were tested for equivalency with regard to difficulty, as were the 20 to 30 Standard Progress Monitor RAPs within each grade. The Benchmark Standard RAPs should be administered three times per school year (1 passage administered in fall, 1 in winter, and 1 in spring), and the Standard Progress Monitor RAPs should be administered as needed throughout the school year.

Previous Version: No information available.

NCEE or REL Study Use:[6] Closing the Reading Gap

[1] In addition to the downloadable format, the Oral Reading Fluency is available for purchase in printed form and as an AIMSweb system (see publisher's web site for more details).

[2] The trainings are often used to "train the trainer" on AIMSweb measures with Response to Intervention (RTI) components in cases where RTI certification is required. The trainings are not required for administration of the Oral Reading Fluency.

[3] Costs apply to materials for a sample of 30 or more students.

[4] The reliability rating refers to internal consistency reliability, which was required for direct assessments (see Appendix A). See the reliability section in the profile narrative for information available on test-retest and alternate form reliability.

[5] Passages for grade 8 at the extremes were removed because of low alternate form reliability. As a result, the mean WRC was lower and the standard deviation greater than for grade 7 passages. According to the authors, the final set of passages for grade 8 is considered to yield more reliable information even though the overall mean is lower than for the grade 7 passages.

[6] See Table F.1 for web address.

References:

Howe, Kathryn B., and Michelle M. Shinn. "Standard Reading Assessment Passages (RAPs) for Use in General Outcome Measurements: A Manual Describing Development and Technical Features." Eden Prairie, MN: Edformation, Inc., 2002.

Kennedy, Jillyan. "AIMSweb Training Presentation: Administration and Scoring of Reading Curriculum-Based Measurement (R-CBM) for Use with AIMSweb." Available at [http://www.aimsweb.com/support-training/training/training-materials/]. n.d.

Shinn, Mark R., and Michelle M. Shinn. *AIMSweb Training Workbook: Administration and Scoring of Reading Curriculum-Based Measurement for Use in General Outcome Measurement.* Eden Prairie, MN: Edformation, Inc., 2002.

Shinn, Mark R., and Michelle M. Shinn. *Illustration 1 and 2.* Eden Prairie, MN: Edformation, Inc., 2002.

Shinn, Mark R., and Michelle M. Shinn. *Practice Exercise 1–8.* Eden Prairie, MN: Edformation, Inc., 2002.

Torgesen, Joseph, Allen Schirm, Laura Castner, Sonya Vartivarian, Wendy Mansfield, David Myers, Fran Stancavage, Donna Durno, Rosanne Javorsky, and Cinthia Haan. "National Assessment of Title I Final Report. Volume II: Closing the Reading Gap: Findings from a Randomized Trial of Four Reading Interventions for Striving Readers." (NCEE 2008-4013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2007.

White, Sheida. *Listening to Children Read Aloud: Oral Fluency*. Washington, DC: National Center for Education Statistics, U.S. Department of Education, 1995.

ALGEBRA END-OF-COURSE ASSESSMENT, 2006[1]

Authors: Educational Testing Service (ETS)		**Type of Assessment:** Group-administered assessment **Domain:** Mathematics
Publisher: Educational Testing Service 866-387-5327 http://ww.ets.org/algebra		**Grade/Age Range:** Grade 6 through 12 **Administration Interval:** At the end of a first-year algebra course, but two sections have been administered all at once or separately (fall and spring)
Material, Training, and Scoring Costs: Algebra End-of-Course Assessment (includes four assessment booklets, assessment answer sheets, and instructions on the use of calculators): Not specified Instructional Data Management System (IDMS): Not specified		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2 (bachelor's degree with coursework in measurement and domain) **Personnel for Administration:** Self- or computer-administered; computer-scored **Training for Administration:** Basic test timing and proctoring
Languages: English		**Alternate Forms:** Four 50-item assessments
Representativeness of Norming Sample: None described		**Summary** **Initial Material Cost:** To be determined upon negotiation with the publisher **Time to Administer:** 40 minutes for each section of assessment **Ease of Administration and Scoring:** 2 (self-administered or administered and scored by someone with basic clerical skills) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available[2] **Construct/Concurrent Validity:** Not available[2] **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Algebra End-of-Course Assessment is a group-administered assessment for grade 6 through 12 students taking first-year algebra. Administration may be paper-and-pencil or online. The use of calculators is not required and varies by district. The two sections of the assessment involve 25 multiple-choice items each (50 items per assessment). Administration of one section takes 40 minutes.

Other Languages: None.

Uses of Information: The publisher states that the Algebra End-of-Course Assessment assesses algebra achievement in order to determine students' ability to use algebraic thinking, identify areas for improvement, and inform and improve teaching strategies.

Methods of Scoring: Using the Instructional Data Management System (IDMS), students answer the multiple-choice questions online; the system then scores the questions electronically. Assessors may also access the assessment online and print copies for paper-and-pencil administration and then scan for scoring with the IDMS. Manual scoring is required for hard copy assessments from ETS. Results are reported as scale scores (metric not specified) and the percent of correct items.

Interpretability: Scores are available in aggregated form or disaggregated by gender, race/ethnicity, grade, and course level by using the IDMS. Item analysis (percentage of students selecting each option for each item) and student ranking by performance level are also available through the IDMS.

Reliability:
(1) Internal consistency reliability: The reliability estimate for scores from a customized form mentioned in *Effectiveness of Reading and Mathematics Software Products: Findings From Two Student Cohorts* (Campuzano et al. 2009) was 0.87 based on 20,506 students.[3] The reliability coefficient for scores from each section of the assessment was similar (Campuzano et al. 2009).
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: No information available.
(4) Inter-rater reliability: No information available.

Validity Evidence:
ETS developed the Algebra End-of-Course Assessment by using the four components of the Algebra Standard outlined by the National Council of Teachers of Mathematics (NCTM) to include (1) understanding patterns, relations, and functions; (2) using algebraic symbols; (3) using mathematical models; and (4) analyzing change.

Construct/Concurrent validity: No information available.[2]

Predictive validity: No information available.[2]

Bias Analysis: No information available.

Training Support: The administrator's manual developed for *Effectiveness of Reading and Mathematics Software Products: Findings From Two Student Cohorts* (Campuzano et al. 2009) provides guidance on how to proctor the assessment, including information on what materials to bring, such as no. 2 pencils and a stopwatch. The manual also includes a script for administration as well as information on how to respond to student questions.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: Four 50-item assessments are available, although it is not specified whether each section or each assessment serves as an alternate form.

Previous Version: None.

NCEE or REL Study Use:[4] Evaluation of the Effectiveness of Educational Technology Interventions

[1] The measure is currently under revision (personal communication with C. Frech, ETS, January 15, 2009).

[2] Campuzano et al. 2009 cites a technical report for the Algebra End-of-Course Assessment; however, we were unable to obtain the report from the publisher.

[3] For the NCEE study, ETS separated the items into two balanced halves with equal levels of difficulty such that one could be administered in fall and the other in spring.

[4] See Table F.1 for web address.

References:

Campuzano, Larissa, Mark Dynarski, Roberto Agodini, and Kristina Rall. "Effectiveness of Reading and Mathematics Software Products: Findings From Two Student Cohorts." (NCEE 2009-4041). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

Dynarski, Mark, Roberto Agodini, Sheila Heaviside, Tim Novak, Nancy Carey, Larissa Campuzano, Barbara Means, Robert Murphy, William Penuel, Hal Javitz, Deborah Emery, and Willow Sussex. "Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort." (NCEE 2007-4005). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2007.

Educational Testing Service. "Algebra End-of-Course Assessment." Available at [http://www.ets.org/Media/Tests/Algebra_End_of_Course_Assessment/pdf/algebra_brochure_eoc.pdf]. 2006.

Educational Testing Service. "Algebra End-of-Course Assessment." Available at [http://www.ets.org/algebra]. 2008.

Educational Testing Service. Algebra Assessment: Research Study, Administrator's Manual. Prepared for Mathematica Policy Research, Princeton, NJ: Educational Testing Service, 2005.

ASSESSING TEACHER LEARNING ABOUT SCIENCE TEACHING (ATLAST)
TEST OF FORCE AND MOTION, 2008

Authors: P. Sean Smith, Iris R. Weiss, Eric R. Banilower, Melanie J. Taylor, and Kimberley D. Wood		**Type of Assessment:** Group-administered assessment **Domain:** Student assessment: science; teacher assessment: subject knowledge (science) and pedagogical content knowledge (science)
Publisher: Horizon Research, Inc. 919-489-1725 http://www.horizon-research.com		**Grade/Age Range:** Middle school students/teachers **Administration Interval:** None described
Material, Training, and Scoring Costs: The teacher and student assessments, answer sheets, answer keys, and domain overview are available at no charge upon request from the developer's web site.[1]		**Personnel and Training Requirements**[1] **Credentials Required for Use:** No special qualifications required **Personnel for Administration:** Test publisher or computer scoring program required[1] **Training for Administration:** None No special qualifications are required to administer the assessment; however, only the developer may score and analyze the results.
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: None described		**Summary** **Initial Material Cost:** 1 (under $100)[1] **Time to Administer:** 30 to 45 minutes **Ease of Administration and Scoring:** 4[1] (administered or scored by a clinician or specialist) **Reliability:** 3[2] (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Not available[3] **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The ATLAST—Test of Force and Motion features group-administered assessments for middle school students and teachers.[4] The student assessment measures students' knowledge and understanding of concepts related to Newton's First Law of Motion and consists of 27 multiple-choice questions covering six concepts within the force and motion construct. The student assessment is available only in hard copy (Horizon Research 2008). The teacher assessment contains three types of questions that measure teachers' knowledge of Newton's First Law of Motion, their ability to analyze student thinking, and their capacity to make instructional decisions to facilitate students' comprehension of the concept (Smith 2006b). The teacher assessment has 29 multiple-choice questions across nine concepts. The teacher assessment may be administered with test booklets or electronically. Both assessments include distracter questions based on common physics misconceptions. Each assessment takes approximately 30 to 45 minutes to complete.

Other Languages: None.

Uses of Information: The ATLAST—Test of Force and Motion measures students' and teachers' growth in their knowledge of concepts related to Newton's First Law of Motion. The assessments may also assess how teacher knowledge affects instruction and student knowledge. The developers state that the assessments should not be used to evaluate students or teachers.

Methods of Scoring: The developers must score both the student and teacher assessments. They will provide the following information: percent of respondents choosing each answer, percent of respondents answering each item correctly, group mean score, and significance testing and computation of effect sizes for pre- and post-test group mean scores (Horizon Research 2008).

Interpretability: The developer provides aggregate analyses for the student and teacher assessments. Teachers may use the results of the student assessment to measure class growth in content area knowledge. Statistically significant gains would suggest that instruction succeeded in imparting the necessary information. Results from the teacher assessments may be used to evaluate knowledge growth following an intervention, such as professional development.

Reliability:
(1) Internal consistency reliability: The developers assessed the internal consistency of scores from both the teacher and student assessments with versions of the assessments containing 25 items. They also conducted a latent dimensionality analysis for the teacher assessment and reported one dimension (P.S. Smith, personal communication, October 17, 2008). Based on IRT results, an internal consistency reliability coefficient for the teacher assessment scores was 0.87 (P.S. Smith, personal communication, October 17, 2008, and October 29, 2008). The reliability coefficient for the student assessment scores was 0.86 (P.S. Smith, personal communication, November 7, 2008).
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: No information available.

Validity Evidence:
The ATLAST—Test of Force and Motion is based on the motion benchmark established by the American Association for the Advancement of Science. The overall construct was divided into 10 concepts and reviewed by a panel of experts. The developers conducted a literature review to identify misconceptions related to the construct on which distracter questions were based. Cognitive interviews with students and teachers also informed development of the test items.

The developers piloted 35 questions for the student assessment with a sample of 2,000 middle school students in spring 2004, and Item Response Theory (IRT) analyses were conducted. They field tested 48 items, divided between two forms with 16 common items, with 5,000 students in fall 2004. (The final version of the assessment contains 27 items on one form.) Using factor and cluster analyses, the developers conducted an investigation of the latent dimensionality for the student assessment, resulting in the identification of two factors: general knowledge of the overall construct and misconceptions related to the concept that constant net force results in constant acceleration. The internal consistency reliability estimate for scores for the latter factor was below 0.40 and was thus dropped in favor of focusing on the more general construct.

The developers piloted 65 multiple-choice questions for the teacher assessment with 1,500 middle and high school physics/physical science teachers in 2005. They conducted IRT analyses, and 33 questions were retained. The developers then administered these questions to 750 teachers; based on IRT results, 25 questions were selected for the final assessment. A panel of experts reviewed and approved the questions.

Construct/Concurrent validity: For student assessment items with difficulties between -2 and +1.5, the test information function demonstrated values above 2 and was highest around the mean item difficulty (value of approximately 6). The item discrimination parameters are greater than 0.40 for all items except the two most difficult items, which are greater than 0.30 (Smith and Banilower 2006).

For teacher assessment items with difficulties between -3 and +2.5, the test information function showed values above 2 and highest at a difficulty of 1 with a value of 8 and a value of 7 at the mean item difficulty of 0 (P.S. Smith, personal communication, November 7, 2008).

To examine associations between student and teacher knowledge, 60 grade 9 teachers completed the ATLAST teacher assessment in summer 2006. Twenty-five of those teachers subsequently administered the student ATLAST to their students before and after a unit on force and motion. A three-level hierarchical linear model (HLM) analysis (scores across time within students within teachers) demonstrated that teachers' scores positively predicted change in students' scores. That is, the students of teachers with higher scores achieved greater gains on the student assessment (P.S. Smith, personal communication, October 17, 2008).

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: When the project's grant funding expires at the end of 2009, assessors will be required to attend a one-day certification workshop that will include training on how to analyze the results. Participants will have to pay for their own travel but will not pay a fee for

attendance. The developers plan to offer online certification in the future (E. Dyer, personal communication, January 29, 2009).

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: None.

Previous Version: The reported reliability and validity information is based on versions of the assessments containing 25 items. The 2008 version of the teacher assessment contains 29 items; the student version has 27 items.

NCEE or REL Study Use:[5] Impact of the Understanding Science Professional Development Model on Science Achievement of English Language Learner Students

[1] When the project's grant funding expires at the end of 2009, the developers will no longer analyze assessment results. Instead, assessors will be required to attend a one-day certification workshop that will include training on how to analyze the results. Participants will pay for their own travel but will not pay a fee for attendance. The developers plan to offer online certification in the future (E. Dyer, personal communication, January 29, 2009).

[2] The reported reliability and validity information is based on versions of the assessments containing 25 items. The 2008 version of the teacher assessment contains 29 items; the student version has 27 items.

[3] Developers reported in a personal communication that they have conducted some analyses of the convergent validity of scores of grade 9 teachers (see Validity Evidence), but they have not published the information.

[4] The ATLAST project at Horizon Research, Inc. developed the assessments. ATLAST is funded by the National Science Foundation under grant number HER-0335328. The project has also developed tests of plate tectonics and flow of matter and energy.

[5] See Table F.1 for web address.

References:

Horizon Research, Inc. "ATLAST: Using the ATLAST Assessments." Available at [http://www.horizon-research.com/atlast/forms.php]. 2008.

Smith, P. Sean, and Eric R. Banilower. "Measuring Middle Grades Students' Understanding of Force and Motion Concepts: Insights into the Structure of Student Ideas." Paper presented at the annual meeting of the National Association for Research in Science, San Francisco, April 2006.

Smith, P. Sean, Iris R. Weiss, Eric R. Banilower, Melanie J. Taylor, and Kimberley D. Wood. "ATLAST Student Assessment: Force and Motion." Chapel Hill, NC: Horizon Research, Inc., 2008.

Smith, P. Sean, Iris R. Weiss, Eric R. Banilower, Melanie J. Taylor, and Kimberley D. Wood. "ATLAST Teacher Assessment: Force and Motion." Chapel Hill, NC: Horizon Research, Inc., 2008.

Smith, P. Sean. "Evaluating Professional Development: New Tool for Assessing Impacts on Teacher Knowledge for Science Teaching." Paper presented at the regional conference of the Mathematics and Science Partnership Program, Orlando, February 16, 2006.

BAYLEY SCALES OF INFANT AND TODDLER DEVELOPMENT, THIRD EDITION (BAYLEY-III), 2005

Authors: Nancy Bayley		**Type of Assessment:** Individually administered adaptive assessment (Cognitive, Language, and Motor scales) and parent report (Social-Emotional and Adaptive Behavior scales) **Domains:**[1] Language arts/language proficiency, cognitive, social-emotional, motor skills, adaptive behavior
Publisher: Pearson Education, Inc. 800-211-8378 http://www.pearsonassess.com		**Grade/Age Range:** 1 to 42 months **Administration Interval:** 3-month interval for children under 12 months of age; 6-month interval for older children
Material, Training, and Scoring Costs: Comprehensive Kit (includes administration and technical manuals, 25 each of Cognitive, Language, and Motor Record Forms; Stimulus Book, Picture Book, manipulative set, 25 Social-Emotional and Adaptive Behavior Questionnaires, 25 Caregiver Report Forms, PDA Administrative Assistant and an administration video): $995		**Personnel and Training Requirements** **Credentials Required for Use:** Level 3 (licensure or state certification, doctorate) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours)
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample:[2] The Cognitive, Language, and Motor scales norming sample was a national, stratified sample of 1,700 children age 1 to 42 months. Based on the 2000 U.S. Census Bureau's Current Population Survey, the sample was stratified along the following variables: gender, region, race/ethnicity, and parent education. The sample consisted of 17 age groups, ranging from 1- to 4-month intervals, with 100 children per group. The children were recruited from health clinics, hospitals, child development centers, churches, and other community organizations and identified by professional recruiters. The initial sample was restricted to typically developing children; then, a subgroup of children with special needs who participated in test development trials (about 10 percent of the sample) was included.		**Summary** **Initial Material Cost:** 4 (>$500) **Time to Administer:** 50 to 90 minutes, depending on age **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Available[3] **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 3 (normed within the past 10 years and nationally representative)

NARRATIVE

Description: The Bayley-III is an individually administered adaptive assessment that measures the developmental functioning of children age 1 to 42 months. The measure presents children with situations and tasks designed to produce an observable set of behavioral responses. It consists of the following scales completed by the assessor: Cognitive Scale, Language Composite Scale with Receptive and Expressive Language subscales, and Motor Composite Scale with Fine- and Gross-Motor subscales. The Bayley-III also includes a Social-Emotional Scale and an Adaptive Behavior Scale for completion by the child's parent or primary caregiver. The assessor completes a Behavior Observation Inventory at the conclusion of the assessment and reviews it with the child's caregiver to determine if the child's behavior during the testing session was typical of the child's usual behavior. Average administration time is 50 minutes for children who are 12 months old or younger and 90 minutes for children age 13 months and older.

The Cognitive Scale assesses development in areas such as concept formation, child play, and number concepts and counting. The Language Composite Scale measures skills such as sound discrimination, word comprehension and production, and imitation. The Motor Composite Scale includes items that assess the quality of movement, sensory integration, and perceptual-motor integration. The number of items per scale ranges from 48 for the Language Composite Scale to 91 for the Cognitive Scale, with varying numbers of questions per item. The items are arranged in order of increasing difficulty, and floor and ceiling rules determine the items administered to each child. The assessor begins at an entry point based on age. The child must respond correctly to the first three items in order to continue. If the child does not respond correctly to the three items, the assessor goes to the entry point for the previous age group to establish the basal or floor. The assessor then administers the item sets until the child gives five incorrect responses in a row, thereby establishing his or her ceiling or highest correct item set.

The Social-Emotional Scale comprises items assessing social-emotional competence and sensory processing; it is based on the Greenspan Social-Emotional Growth Chart: A Screening Questionnaire for Infants and Young Children. The assessment measures functional emotional milestones. The Adaptive Behavior Scale, based on the Adaptive Behavior Assessment System-Second Edition (ABAS-II), assesses the attainment of adaptive behavior skills necessary for infants' and young children's development of independence.

For ongoing developmental screening, the Bayley Scales of Infant and Toddler Development, Third Edition Screening Test, is available. In a 15- to 25-minute administration period, the screener assesses cognitive, language, and motor development. In addition, the Bayley Short Form-Research Edition (BSF-R), based on the Bayley Scale of Infant Development-Second Edition (BSID-II), was developed for the Early Childhood Longitudinal Study-Birth Cohort to provide an assessment that is less time-consuming than the full BSID-II but sufficiently comprehensive to capture adequately the development of infants and young children (Berry et al. 2004).

Other Languages: None.

Uses of Information: The Bayley-III is used to identify areas of relative impairment or delay, develop curricula for interventions, and assess the outcome of such interventions. It is not a

diagnostic tool but rather indicates areas that might require further evaluation. The scales should not be used to assess a child's deficit in a specific skill area or to make a norm-referenced interpretation of scores for children with severe sensory or physical impairments. In addition, although some of the measure's items are similar to items on tests of school-age abilities, the Bayley-III is not an intelligence test.

Methods of Scoring: For the Cognitive, Language, and Motor scales, items are scored as correct (1) or incorrect (0) depending on whether the child displayed the indicated action. The administration manual provides detailed scoring instructions for each item. The raw score is the sum of the child's correct points. All items below the basal are scored as correct. The types of scores available for each scale and subscale vary; in general, by using the tables provided in the manual, the assessor may obtain scaled scores, composite scores, percentile ranks, confidence intervals, developmental age equivalents, and growth scores. Individuals without a Level 3 credential may score the assessment under supervision. Scoring software may be purchased from the publisher.

The Social-Emotional Scale uses a six-point frequency rating (can't tell, none of the time, some of the time, half of the time, most of the time, all of the time). The raw score is the sum of the behavior frequencies. The total for the first eight items provides a sensory processing score. The Adaptive Behavior Scale uses a four-point frequency rating (is not able, never when needed, sometimes when needed, always when needed) and provides the following scores: a subscale score for each of the 10 skill areas; 3 domain area scores for the Conceptual, Social, and Practical domains; and a General Adaptive Composite score that is a sum of each child's scores across the skill areas. Using the tables provided in the manual, the assessor may convert raw scores from both scales into scaled scores, composite scores, and percentile ranks and determine confidence intervals.

Interpretability: Only persons with formal training in test administration should interpret the results of the Bayley-III. The technical manual provides detailed information on how to interpret the scores. Norms are available by age groups of varying intervals (e.g., 10 days to 3 months) to facilitate norm-referenced interpretation of performance during the period of infant and toddler development. The Behavior Observation Inventory provides qualitative information to facilitate interpretation of the child's performance and intervention planning.

Reliability:
(1) Internal consistency reliability: The developers assessed the internal consistency of raw scores from the Cognitive, Language, and Motor scales by using the split-half method based on the entire norming sample. The reliability coefficients for scores on the Cognitive Scale ranged from 0.79 to 0.97 across 17 age groups. The coefficients for scores on the Language Composite Scale ranged from 0.82 to 0.98 and, across its subscales, from 0.71 to 0.97. For raw scores on the Motor Composite Scale, the coefficients ranged from 0.86 to 0.96 and, across its subscales, from 0.72 to 0.95.

The developers also estimated the internal consistency of scores from the Cognitive, Language, and Motor scales with a clinical population of 668 children. For scores on the Cognitive Scale, coefficients ranged from 0.90 to 0.99 across 9 age groups. For the Language Composite Scale, reliability coefficients for scores from the Receptive and Expressive Communication subscales

ranged from 0.74 to 0.99. For the Motor Composite Scale, coefficients for scores from the Fine- and Gross-Motor subscales ranged from 0.84 to 0.99.
For the Social-Emotional and Adaptive Behavior scales, the internal consistency for raw scores comes from the original measures' manuals (see Description) as reported by the Bayley-III developers. For scores from the Greenspan Social-Emotional Growth Chart, reliability coefficients ranged from 0.83 to 0.94 based on the social-emotional items across 8 age groups and from 0.76 to 0.91 based on the sensory processing items. For scores from the ABAS-II, reliability coefficients ranged from 0.86 to 0.98 for the total score (i.e., General Adaptive Composite) across 10 age groups and, for subscale scores, from 0.70 to 0.96 for the Conceptual domain, from 0.81 to 0.94 for the Social domain, and from 0.82 to 0.97 for the Practical domain (except for the scores of children 0 to 3 months, 0.65). For a sample of 246 children with developmental delays, motor impairments, language disorders, and biological risk factors, coefficients for the ABAS-II total score ranged from 0.97 to 0.99 while coefficients for scores from subscales ranged from 0.90 to 0.99.
(2) Test-retest reliability: The developers assessed the test-retest reliability for scores from the Cognitive, Language, and Motor scales by using a sample of 197 children from the norming sample, age 2 to 42 months. The test-retest interval between administrations ranged from 2 to 15 days (mean = 6). The correlations for scores on the Cognitive Scale ranged from 0.75 to 0.86 across four age groups. The correlations for scores on the Language Composite Scale ranged from 0.69 to 0.87 and, for its subscales, from 0.63 to 0.84. The correlations for scores on the Motor Composite Scale ranged from 0.79 to 0.84 and, for its subscales, from 0.73 to 0.86. The developers also present correlations corrected for the variability of the standardization sample.

No test-retest reliability information is reported for scores of the Social-Emotional Scale. The developers assessed test-retest reliability of scores from the Adaptive Behavior Scale by using parent reports for 207 children age 0 to 35 months. The test-retest interval between administrations ranged from 2 days to 5 weeks (mean = 12 days). The correlations for the total score ranged from 0.86 to 0.91 while correlations for subscale scores ranged from 0.81 to 0.90. The developers also present correlations corrected for the variability of the standardization sample.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: Inter-rater reliability was assessed for the Adaptive Behavior Scale, administered to a sample of 56 children age 0 to 71 months, and rated by two parents. The correlation for the total scores was 0.77. Correlations ranged from 0.69 to 0.83 across domains. The developers also present correlations corrected for the variability of the standardization sample.

Validity Evidence:
Expert consultation and a review of the literature guided development of the Cognitive, Language, and Motor scales. In addition, an advisory panel lent guidance throughout the development process. The developers also consulted clinical measurement specialists, and conducted focus groups and surveys. The development of the Bayley-III occurred in several stages with pilot and tryout studies. Information on the content validity of the Greenspan Social-Emotional Growth Chart, on which the Social-Emotional Scale is based, is not provided. The developers of the ABAS-II established its set of daily independent living skills based on legal and professional concepts, standards, and regulations related to special education and developmental disability.

Construct/Concurrent validity: A confirmatory factor analysis using the norming sample data supported a three-factor latent structure for the Cognitive, Language, and Motor scales (based on the root mean square error of approximation). The first factor was the Motor Composite Scale while the Language Composite Scale and Cognitive Scale were the second and third factors. Intercorrelations between subscale and composite scores (e.g., Fine-Motor and Gross-Motor with the Motor Composite) provided another test of the validity of the measure. The correlation between scores on the Language Composite Scale and scores on the Receptive and Expressive Communication subscales was 0.71. Correlations between scores on the Language Composite Scale and scores on the other scales ranged from 0.25 (Social-Emotional) to 0.52 (Cognitive). Correlations between scores on the Motor Composite Scale and scores on the Fine- and Gross-Motor subscales ranged from 0.69 to 0.70. Correlations between scores on the Motor Composite Scale and scores on the other scales ranged from 0.21 (Social-Emotional) to 0.51 (Cognitive). Corrected correlations between the subscales that comprise the composites are provided.

During standardization, the developers compared the Bayley-III scores to scores on the following instruments: Bayley Scale of Infant Development-Second Edition (BSID-II), Weschler Preschool and Primary Scale of Intelligence-Third Edition (WPPSI-III), Preschool Language Scale-Fourth Edition (PLS-4), and Peabody Developmental Motor Scales-Second Edition (PDMS-2). Separate samples were used for each combination of the Bayley-III with one other measure. Samples ranged from about 50 to 100 children age 0 to 42 months (except for the WPPSI-III, which included children age 28 to 42 months), and the test intervals ranged from 0 to 28 days (mean = 5 to 6). The administration order of the assessments was counterbalanced, and the means were compared across the two orders of administration. Correlations were corrected for the variability of the standardization sample.

The correlations between scores on the Bayley-III Cognitive Scale and other cognitive measures were 0.60 for the BSID-II Mental Scale, 0.72 to 0.79 for the WPPSI-III, and 0.57 for the PLS-4 Composite. Correlations between scores on the Bayley-III Language Composite and other language measures were 0.71 for the BSID-II Mental Scale, 0.71 to 0.83 for the WPPSI-III, and 0.66 for the PLS-4 Composite with correlations of 0.62 and 0.68, respectively, between the receptive and expressive communication subtest scores. The correlation between scores on the Bayley-III Motor Composite and the PDMS-II Total Motor Quotient was 0.57 and, between the fine- and gross-motor subtest scores, 0.59. The correlation between the Bayley-III Social-Emotional Scale and the BSID-II Behavior Rating Scale was 0.38.

The ABAS-II (the basis for the Bayley-III Adaptive Behavior Scale) and the Vineland Adaptive Behavior Scale-Interview Edition (VABS) were administered to a sample of 45 typically developing children age 1 to 69 months (mean = 34). The correlation between their scores was 0.70.

The Bayley-III subscales were also correlated with other scales measuring largely different skills. The correlations for the Bayley-III Cognitive Scale scores were 0.40 and 0.38, respectively, with the BSID-II Motor and Behavior Rating scales and 0.45 with the PDMS-II total motor score. The correlations between scores on the Bayley-III Language Composite Scale and the BSID-II Motor and Behavior Rating Scales were 0.47 and 0.37, respectively, and 0.45 with the PDMS-II total motor score. The correlations between scores on the Bayley-III Motor

Composite Scale and the BSID-II Mental and Behavior Rating Scales were 0.44 and 0.31, respectively, 0.55 with the WPPSI-III, and 0.44 with the PLS-4 Composite. The correlations for the Bayley-III Social-Emotional Scale scores were 0.25 and 0.24, respectively, with the BSID-II Mental and Motor Scales scores, 0.43 with the WPPSI-III, 0.23 with the PLS-4 Composite, and 0.25 with the PDMS-II total motor score.

The Bayley-III and the ABAS-II (the basis for the Bayley-III Adaptive Behavior Scale) were administered to a sample of 60 children age 5 to 37 months with a test interval of 0 to 23 days (mean = 4). The correlations between scores on the Bayley-III scales and the ABAS-II General Adaptive Composite ranged from 0.25 to 0.36, except for ABAS-II General Adaptive Composite scores with the Social-Emotional scale, which correlated 0.04. The ABAS-II and the Scales of Independent Behavior-Revised: Early Development Form (SIB-R), a brief 40-item assessment, were administered to a typically developing sample of 34 children age 2 to 23 months (mean = 14); the correlation between scores was 0.18.

The developers assessed the Bayley-III's ability to differentiate (not diagnose) between convenience samples of special populations and a group of typically developing children matched on gender, parent education level, race/ethnicity, and geographic region. The scores on all subscales and composites as compared to children without disabilities were significantly lower for the following groups: children with Down syndrome, children with pervasive development disorder, children with cerebral palsy, children with specific or suspected language impairment, and children at risk for developmental delay. Children with asphyxiation at birth scored significantly lower than children in the control group on all subscale and composite scores except for the Expressive Language subscale. For children with prenatal alcohol exposure, the mean difference between scores was significant for all subscales and composites except for the Motor Composite and its associated subscales. For children small for gestational age, the mean differences were statistically significant for the Language and Motor Composite scores and the Receptive Communication and Gross-Motor subscales. Children born prematurely or with low birth weight evidenced significant differences in the Motor Composite Score and the Fine-Motor subscale.

The developers also examined Social-Emotional scale scores. The percentage of children scoring two or more standard deviations below the mean was greater for the special population groups (4 to 67 percent) than for the control groups (0 to 3 percent). In terms of the ABAS-II (the basis for the Bayley-III Adaptive Behavior Scale), the developers found significant differences for all skill areas and domains between the following groups of children and their matched controls: children with developmental delays, children with biological risk factors, children with motor and physical impairments, and children with receptive and/or expressive language disorders.

Predictive validity: Predictive validity analyses were conducted with the original Bayley Scales of Infant Development (BSID 1969), which contained Mental and Motor scales. The author concludes, ". . .the BSID at the scale level is generally not as predictive of later intellectual, language, or achievement performance as are specified subscales; however, the later in the preschool period (i.e., beyond two years) the BSID scores are obtained, the more predictive they are of later childhood functioning" (Bayley 1993).

Bias Analysis: During the test development process, potential item bias was assessed through expert review and statistical analyses, resulting in the deletion of 30 items. The developers tested an additional 120 Black and Hispanic children to ensure adequate sample sizes for conducting analyses of these groups. Analysis for differential item functioning (DIF) was conducted by using the Mantel-Haenszel method. The developers do not provide details on the DIF analysis results.

Training Support: The Bayley-III may be purchased only by individuals highly trained in test administration and interpretation as evidenced by a doctorate degree, certification, or licensure. The administration manual provides detailed information on how to administer and score the assessment. A training video and an interactive administration and scoring DVD are available for purchase.

Adaptations/Special Instructions for Individuals with Disabilities: The administration manual provides information on modifications that may be made to accommodate the assessment of children with disabilities. For example, with respect to assessing children with hearing impairments, the manual notes that light sources should be placed in front of the assessor to reduce glare. It also states that children may be prompted by slightly moving the manipulative to draw a child's attention.

Alternate Forms: None.

Previous Version: The Bayley-III updates the Bayley Scale for Infant Development-Second Edition (BSID-II) published in 1993. It expands content coverage by splitting the Mental Scale into the Cognitive and Language scales and replacing the BSID-II Behavior Rating Scale with the Social-Emotional and Adaptive Behavior scales. In addition, it adds new items to the Cognitive, Language, and Motor scales and includes the Behavior Observation Inventory to gauge whether a child's behavior during the assessment was representative of the child's typical conduct. The updated measure extends the floor and ceiling for each scale by including, respectively, gifted children and children with or at risk of developmental challenges. Stimulus materials were updated and printed in color, and procedures were modified to increase ease of administration.

NCEE or REL Study Use:[4] Program for Infant and Toddler Caregivers (PITC) (REL-West)

[1] This measure assesses infant and toddler development; most domains outlined for student school outcomes do not apply.

[2] The Social-Emotional Scale was normed during the Bayley-III test development phase in spring 2003. The norming sample was a stratified sample of 456 U.S. children age 1 to 42 months. Based on the 2000 U.S. Census Bureau's Current Population Survey, the sample was stratified along the following variables: Census region, race/ethnicity, and parent education. The sample consisted of eight age groups of approximately equal numbers of males and females, with 50 to 89 children per group. Children were excluded from the sample if they did not speak or understand English; had hearing or visual impairments; had developmental risk factors based on social, socioeconomic status, or parent education factors; or were taking medication that could affect test performance.

The standardization, reliability, and validity evidence for the Adaptive Behavior Scale are based on the norming sample data of the ABAS-II, which included a stratified sample of 1,350 children from 0 to 71 months divided into 13 groups of 3- to 5-month intervals of 100 children each, except for the oldest age group, which had 150 children. There were equal numbers of males and females in each age group. The sample was stratified by race/ethnicity and parent education level based on the October 2000 U.S. Census Bureau's Current Population Survey, and efforts were made to ensure the sample was geographically representative. Children with special needs comprised 2.88 percent of the sample.

[3] The predictive validity information is based on the original Bayley Scales of Infant Development (BSID 1969).

[4] See Table F.1 for web address.

References:

Albers, Craig A., and Adam J. Grieve. "Test Review: Bayley Scales of Infant and Toddler Development-Third Edition." *Journal of Psychoeducational Assessment*, vol. 25, no. 2, 2007, pp. 180-190.

Bayley, Nancy. *Bayley Scales of Infant Development-Second Edition Manual*. San Antonio, TX: PsychCorp, 1993.

Bayley, Nancy. *Bayley Scales of Infant and Toddler Development-Third Edition: Administration Manual*. San Antonio, TX: PsychCorp, 2006.

Bayley, Nancy. *Bayley Scales of Infant and Toddler Development-Third Edition: Technical Manual*. San Antonio, TX: PsychCorp, 2006.

Berry, Daniel J., Lisa J. Bridges, and Martha J. Zaslow. "Early Childhood Measures Profiles." Washington, DC: Child Trends, 2004.

Kisker, Ellen E., Kimberly Boller, Charles Nagatoshi, Christine Sciarrino, Vinita Jethwani, Teresa Zavitsky, Melissa Ford, and John M. Love. "Resources for Measuring Services and Outcomes in Head Start Programs Serving Infants and Toddlers." Washington, DC: Mathematica Policy Research, 2003.

Pearson Assessments. "Bayley-III Enhanced Administration/Scoring Resource Interactive DVD." Available at [http://pearsonassess.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=015-8027-23X&Mode=detail&Leaf=accessory&dsrc=015-8027-612#ISBN2]. 2005.

Pearson Assessments. "Bayley-III Fundamental Administration Video." Available at [http://pearsonassess.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=015-8027-23X&Mode=detail&Leaf=accessory&dsrc=015-8027-604#ISBN2]. 2005.

Pearson Assessments. "Bayley-III Scoring Assistant® and PDA Administration Software." Available at [http://pearsonassess.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=015-8027-23X&Mode=detail&Leaf=software]. 2005.

Tobin, Renee M., and Kathryn E. Hoff. "Review of the Bayley Scales of Infant and Toddler Development-Third Edition." In *The Seventeenth Mental Measurements Yearbook*, edited by Kurt F. Geisinger, Robert A. Spies, Janet F. Carlson, and Barbara S. Plake. Lincoln, NE: The Buros Institute of Mental Measurements, 2007.

Venn, John J. "Review of the Bayley Scales of Infant and Toddler Development-Third Edition." In *The Seventeenth Mental Measurements Yearbook*, edited by Kurt F. Geisinger, Robert A. Spies, Janet F. Carlson, and Barbara S. Plake. Lincoln, NE: The Buros Institute of Mental Measurements, 2007.

DYNAMIC INDICATORS OF BASIC EARLY LITERACY SKILLS (DIBELS)
SIXTH EDITION, 2007

Authors: Roland H. Good and Ruth A. Kaminski		**Type of Assessment:** Individual assessment **Domain:** Reading (phonological awareness, letter recognition and naming, vocabulary, decoding, phonics, reading fluency, various comprehension skills)
Publisher: University of Oregon Center on Teaching and Learning (free downloadable materials) 888-497-4290 https://dibels.uoregon.edu Sopris West Educational Services (print materials) 800-547-6747 http://www.sopriswest.com Wireless Generation (handheld computer software) 800-823-1969 http://www.wirelessgeneration.com		**Grade/Age Range:** Kindergarten through grade 6[1] **Administration Interval:** Three times per school year for Benchmark Assessments; as often as desired for Progress Monitoring Assessments
Material, Training, and Scoring Costs: Free, reproducible downloads of materials available at https://dibels.uoregon.edu Print materials: Classroom sets (DIBELS Administration and Scoring Guide, 25 Benchmark Assessment sheets, 6 Progress Monitoring Scoring Booklets, and Student Materials): $72.49 for kindergarten through grade 3 sets, $57.49 for grade 4 through 6 sets (separate set required for each grade)		**Personnel and Training Requirements** **Credentials Required for Use:** Level 1 (training or supervised experience with measurement) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (> 2 hours)
Languages: English, Spanish—see Indicadores Dinámicos del Éxito en la Lectura (IDEL) profile		**Alternate Forms:** For progress monitoring, 20 forms (passages, tasks, probes) provided for each subtest
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (under $100) **Time to Administer:** About 10 to 15 minutes (1 to 3 minutes per subtest) **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 1[2] (none described) **Predictive Validity:** Available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: Dynamic Indicators of Basic Early Literacy Skills (DIBELS) encompass a set of seven individually administered, standardized screening procedures and measures for assessing elementary school students' acquisition of early literacy and reading skills. DIBELS provides repeated, direct assessment in five areas: (1) phonemic awareness, (2) phonics, (3) fluency, (4) comprehension, and (5) vocabulary. Phonemic awareness subtests include Initial Sound Fluency (ISF) and Phoneme Segmentation Fluency (PSF), and phonics subtests include Nonsense Word Fluency (NWF) and Letter Naming Fluency (LNF). The Oral Reading Fluency (ORF), Retell Fluency (RTF), and Word Use Fluency (WUF) subtests assess fluency, comprehension, and vocabulary, respectively. The ISF subtest is administered to kindergarten students only. PSF and LNF are appropriate for kindergarten and grade 1 students, and NWF may be administered to kindergarten and grade 1 and 2 students. WUF may be administered to kindergarten through grade 3 students, and ORF and RTF may be administered to students in grade 1 through 6. The subtests assess students' performance on key indicators of these skills but do not comprehensively assess mastery of skill areas. Although the authors state that the DIBELS measures are appropriate for students in kindergarten through grade 6, empirically derived cut scores are available only for students in kindergarten through grade 3.

The DIBELS framework includes Benchmark Assessments, which comprise grade-appropriate selections of subtests (described above) with total administration time less than 15 minutes, and Progress Monitoring Assessments, which assess student progress and intervention effectiveness between Benchmark Assessments. Educators may administer the Progress Monitoring Assessments as frequently as necessary, using up to 20 alternate assessment tasks to prevent practice effects.

Other Languages: Indicadores Dinámicos del Éxito en la Lectura (IDEL) assesses the basic early literacy skills of students learning to read in Spanish (see IDEL profile in the current compendium). The IDEL measures are not a translation of the DIBELS, though the measures are based on similar theory and research about how students learn to read in alphabetic languages such as English and Spanish.

Uses of Information: DIBELS assessment data may be used to identify students who are experiencing reading difficulties and therefore are at risk for ongoing reading problems. The data may be used to plan, evaluate, or modify instructional support for identified students and to monitor outcomes. In addition, the data may be aggregated to track and compare the progress of groups of students and monitor the effectiveness of reading instruction and interventions in classrooms, schools, and districts.

Methods of Scoring: Assessors must hand-score DIBELS assessments. The Administration and Scoring Guide (Good and Kaminski 2002) provides guidelines for scoring responses as correct or incorrect. Scorers calculate raw scores based on the number of correct responses and compare them with decision rules (based on cut points) linked to descriptors of students' need for support. Alternatively, schools or districts that upload local sample data into the DIBELS Data System may request local percentile ranks, custom reports, and summaries based on their own data.

Interpretability: DIBELS benchmark goals and related cutoff scores facilitate criterion-referenced interpretation of scores. According to the authors, benchmark goals for each measure and time period indicate the probability of achieving the next benchmark goal. Students whose scores on one or more DIBELS measures fall at or above the benchmark have at least an 80 percent chance of meeting the next benchmark goal (an indicator of appropriate progress). The Administration and Scoring Guide includes categories of students' need for support. Based on subtest scores, assessors may categorize a student's need for support as Benchmark (having met the benchmark goal), Intensive (20 percent or less probability of achieving the next benchmark goal), or Strategic (21 to 50 percent probability of achieving the next benchmark goal). Schools may also examine student performance in comparison to school or district peers. The authors recommend that schools consider student performance in relation to the benchmarks rather than percentiles; the former are predictive of future success.

The DIBELS Data System reports results at the student, class, school, program, and district levels. Data system users may request several types of reports, including individual student profiles, class reports (name, scores, percentiles, instructional status for all students in a class), school and district summary reports (means and proficiency level across the school year for all measures), distribution reports (disaggregated results by school, class, demographics), and district norms. Users may view reports on web pages or download PDF files.

Reliability:[3]
(1) Internal consistency reliability: No information available.
(2) Test-retest reliability: Researchers have reported test-retest reliability coefficients ranging from 0.92 to 0.97 (Good and Kaminski 2002) and 0.94 to 0.98 (Baker et al. 2008) for ORF scores. For NWF scores, Harn et al. (2008) reported a test-retest reliability coefficient of 0.64.
(3) Alternate form reliability: Good et al. (2008) reported median alternate form reliability coefficients for ORF scores ranging from 0.92 to 0.95 for separate samples for grade 1 through 6. The following alternate form reliability coefficients have also been reported: 0.89 to 0.94 for LNF scores (Good et al. 2004; Hintze et al. 2003); 0.61 to 0.86 for ISF scores (Good et al. 2004; Hintze et al. 2003); 0.74 (Good et al. 2004) and 0.97 (Hintze et al. 2003) for PSF scores; 0.83 to 0.94 for NWF scores (Good et al. 2004; Harn et al. 2008; Ritchey 2008; Speece et al. 2003); 0.89 to 0.97 for ORF scores (Baker et al, 2008; Roberts et al. 2005); and 0.57 to 0.90 for RTF scores (Good et al. 2004; Roberts et al. 2005). Kaminski et al. (2004) reported an alternate form reliability coefficient of 0.88 for PSF scores and median alternate form reliability coefficients ranging from 0.52 to 0.71 for the WUF subtest scores.
(4) Inter-rater reliability: No information available.

Validity Evidence:
The benchmark goals and decision rules for instructional recommendations are based on longitudinal predictive information from samples participating in the DIBELS Data System. Cross-year predictive utilities are based on all participating schools during the 2000–2001 and 2001–2002 school years, and within-year predictive utilities are based on all participating schools during the 2001–2002 school year. Sample sizes for most analyses ranged from 32,000 to 34,794, except for the end-of-grade 1 benchmark assessment (N = 6,239). The authors provide little information about the characteristics of the samples and acknowledge that they are non-representative convenience samples.

The DIBELS measures assess the five foundational early reading skill areas identified by the National Reading Panel (2000) as prerequisites to later reading success: (1) phonemic awareness, (2) phonics, (3) fluency, (4) comprehension, and (5) vocabulary. The authors provide minimal information about development of the items and the sample with which the measures were first developed.

Construct/Concurrent validity: Hagan-Burke et al. (2006) examined the factor structure of DIBELS by using data from the NWF, LNF, PSF, and WUF subtests administered to a sample of 202 grade 1 students. Most of the students were White (63.9 percent), but also included Black (27.2 percent), multiracial (4.5 percent), Hispanic (3.5 percent), and Asian (1 percent) students. The sample comprised students from middle- and lower-middle-class families (38.6 percent received free or reduced-price lunch). All four assessments loaded on a single factor, which accounted for 39.5 percent of the variance among correlated indicators.

In studies conducted with kindergarten and grade 1 students, researchers have found positive correlations of varying magnitudes between scores on DIBELS subtests and other assessments of early reading proficiency. Elliot et al. (2001) reported correlations ranging from 0.60 to 0.70 between kindergartners' scores on a modified version of the DIBELS and the Woodcock-Johnson Psycho-Educational Achievement Battery-Revised (WJ-R) Broad Reading and Skills clusters, the Test of Phonological Awareness (TOPA), and teacher ratings of student achievement. Speece et al. (2003) found correlations of 0.71 and 0.75 between grade 1 students' scores on NWF and scores on the WJ-R Letter-Word Identification and Word Attack subtests, respectively. Similarly, Good and Kaminski (2004)[3] reported a median correlation of 0.70 between LNF and Woodcock-Johnson Psycho-Educational Battery readiness cluster scores. They also cited correlations ranging from 0.26 to 0.59 between scores on the ISF, PSF, NWF, and WUF subtests and scores on the Woodcock-Johnson readiness cluster, Test of Language Development-Primary, Third Edition (TOLD-3), and Peabody Picture Vocabulary Test III (PPVT-III). Other studies reported similar results in studies comparing DIBELS subtest scores to scores on other measures of early reading and language proficiency (Hagan-Burke et al. 2006; Hintze et al. 2003; Rouse and Fantuzzo 2006).

Several studies have investigated relationships between student performance on the DIBELS ORF subtest and "high stakes" state reading tests. Studies conducted with students in grade 3 through 5 in Florida (Buck and Torgesen 2003; Roehrig et al. 2008), Colorado (Shaw and Shaw 2002), Ohio (Vander Meer et al. 2005), North Carolina (Barger 2003), Arizona (Wilson 2005), Delaware (Uribe-Zarain 2007), and Pennsylvania (Shapiro et al. 2008) reported correlations ranging from 0.52 to 0.80 between DIBELS ORF scores and scores on state reading proficiency tests.

Rouse and Fantuzzo (2006) reported negative correlations ranging from -0.10 to -0.30 between kindergarten students' scores on the LNF, PSF, and NWF DIBELS subtests and the Disruption and Disconnection dimensions of the Penn Interactive Peer Play Scale (PIPPS). As expected, they also found that measures of positive social interaction and learning behaviors correlated to a lesser degree with DIBELS subtest scores than did cognitive and language measures (coefficients ranged from 0.17 to 0.52). Using multivariate analyses provided further evidence of divergent validity. Hagan-Burke et al. (2006) reported correlations ranging from 0.30 to 0.37 between scores on the WUF subtest (which measures vocabulary) and scores on DIBELS and TOWRE

measures of word reading and phonological awareness. Schilling et al. (2007) correlated NWF and ORF scores with scores on the Iowa Test of Basic Skills Listening subtest, finding correlation coefficients ranging from 0.29 to 0.37.

Reviewers have commented that the DIBELS developers have not clearly documented the validity of the benchmark cut scores, decision rules, and associated instructional classifications (Brunsman 2005; Shanahan 2005). Hintze et al. (2003) investigated the diagnostic accuracy of the DIBELS by using classifications on the Comprehensive Test of Phonological Processing (CTOPP) as a criterion measure. Receiver-operating characteristic curve (ROC) analyses indicated that use of the DIBELS cut scores resulted in high sensitivity and low specificity, suggesting that the scores may result in an inordinate number of false positives (students whose DIBELS scores suggested reading problems but whose CTOPP scores did not). The developers stated that, although such outcomes may be acceptable as long as the DIBELS is used for screening rather than for diagnostic purposes, test users may want to use recalibrated cut scores. With a sample of 32,307 grade 3 students in Florida, Roehrig et al. (2008) also determined that recalibrated ORF risk-level cut scores determined via ROC curve analyses more accurately identified true positives than previously established DIBELS benchmarks.

Predictive validity: Several studies have reported positive predictive validity coefficients of varying magnitude between scores on DIBELS subtests and other assessments of early reading proficiency. Good et al. (2004) reported correlations ranging from 0.28 to 0.69 between kindergarten students' ISF, PSF, and LNF scores (collected in winter to spring) and end-of-grade 1 Woodcock-Johnson Psycho-Educational Battery total reading cluster scores. For grade 1 students, correlations ranged from 0.20 to 0.77 between fall-to-spring PSF, NWF, and LNF scores and end-of-grade 2 Woodcock-Johnson Psycho-Educational Battery total reading cluster scores. Several studies found that elementary students' scores on DIBELS subtests predicted performance on standardized reading and language measures administered 2 to 24 months later. Across the studies, predictive validity coefficients ranged from 0.18 to 0.89 (Baker et al. 2008; Burke et al. as cited in Good et al. 2008; Powell-Smith et al. 2008; Riedel 2007; Ritchey 2008; Rouse and Fantuzzo 2006).

Researchers have also investigated the predictive validity of the DIBELS ORF subtest in relation to state-mandated reading tests. In studies conducted with students in grade 3 through 5 in Florida (Roehrig et al. 2008), Colorado (Shaw and Shaw 2002; Wood 2006), Ohio (Vander Meer et al. 2005), and Pennsylvania (Shapiro et al. 2008), researchers reported correlations ranging from 0.61 to 0.75 between ORF scores and scores on state reading assessments administered later in the same school year or the following school year.

Bias Analysis: Roehrig et al. (2008) conducted logistic regression analyses to examine whether subgroups differed with respect to percentages of students meeting the end-of-year reading comprehension benchmark on the Florida state test (FCAT-SSS) according to ORF risk classification. In separate models, they entered independent variables for a demographic characteristic (i.e., race/ethnicity, socioeconomic status, language status), ORF risk classification, and an interaction term of the demographic predictor with ORF risk classification. The analyses revealed no bias for predicting the level of performance on the state assessment using ORF scores across racial/ethnic, socioeconomic, or language groups. The authors' sample consisted of 35,207 grade 3 demographically diverse students in Florida Reading First schools

during the 2004–2005 school year. Seventy-five percent of students were eligible for free or reduced-price lunch, and 17 percent were students with a disability. Other studies with the DIBELS found similar results (Buck and Torgesen 2003; Wilson 2005).

Training Support: The authors state that "most educational personnel" can self-train from materials available on the DIBELS web site (https://dibels.uoregon.edu). Sopris West Educational Services publishes printed and video/DVD training and implementation materials (see References). DIBELS users may also participate in training workshops offered by either Dynamic Measurement Group (http://www.dynamicmeasurement.org) or Sopris West. Founded by the DIBELS authors, Dynamic Measurement Group offers on-site training, one- and two-day regional workshops for individuals or school-based teams, and a four-day summer training institute led by the authors. The training sessions address DIBELS administration, scoring, interpretation, and data management. Sopris West offers two-day training sessions covering administration, scoring, and planning instruction with DIBELS data.

Adaptations/Special Instructions for Individuals with Disabilities: DIBELS is not appropriate for use with students who (1) are deaf, (2) have fluency-based speech disabilities, (3) are learning to read in a language other than English, or (4) have severe disabilities. The measure may be used for students with other types of identified disabilities; in some cases, educators may need to adjust goals, timelines, and materials (e.g., out-of-grade level testing).

The DIBELS Administration and Scoring Guide (Good and Kaminski 2002) lists several accommodations that assessors may use when testing students "for whom a standard administration may not provide an accurate estimate of their skills . . ." (p. 44). With visually impaired students, assessors may use a large-print version or Braille student stimulus materials or visually enhanced stimulus materials (e.g., enhanced size, lighting, alternate print fonts, color printing). For students who experience difficulties with the assessment task, assessors may retest the student under different conditions or use approved accommodations for altering subtest instructions. Additionally, the student may be tested by different assessors in such cases.

Alternate Forms: For progress monitoring, alternate forms (tasks, passages, prompts) are available for all subtests except for LNF, which is a risk indicator and is not monitored over time. Test users may choose from 20 alternate forms for ISF, PSF, and WUF, 20 alternate forms per grade for ORF and RTF (which use the same forms), and two sets of 20 alternate forms (one for kindergarten and grade 1 and one for grades 2 and 3) for WUF. The authors state that, based on similar readability levels and correlations across forms, alternate forms for DIBELS subtests are equivalent. Francis et al. (2008), however, found substantial differences in difficulty among six ORF reading passages and recommend that test users correct for difficulty effects by using statistical methods for equating passages.

Previous Version: Before publication of the DIBELS Sixth Edition in 2002, the DIBELS Fifth Edition (Good and Kaminski 2001) was available only by download through the University of Oregon Center on Teaching and Learning. The Sixth Edition includes all new reading passages and prompts and introduced the ORF progress monitoring probes.

NCEE or REL Study Use:[4] The Effects of Opening the World of Learning (OWL) on the Early Literacy Skills of At-Risk Urban Preschool Students

[1] Subtests vary in the appropriate grade range (see Description). Cut scores, decision rules, instructional recommendations are only available for kindergarten through grade 3 students.

[2] The reliability rating refers to internal consistency reliability, which was required for direct assessments (see Appendix A). See the reliability section in the profile narrative for information available on test-retest and alternate form reliability.

[3] The DIBELS Administration and Scoring Guide (Good and Kaminski 2002) cites selected reliability and validity coefficients but does not describe sample characteristics or study designs.

[4] See Table F.1 for web address.

References:

Baker, Scott K., Keith Smolkowski, Rachel Katz, Hank Fien, John R. Seeley, Edward J. Kame'enui, and Carrie T. Beck. "Reading Fluency as a Predictor of Reading Proficiency in Low-Performing, High-Poverty Schools." *School Psychology Review,* vol. 37, no. 1, 2008, pp. 18-37.

Barger, Jeff. "Comparing the DIBELS Oral Fluency Indicator and the North Carolina End of Grade Reading Assessment (Technical Report)." Asheville, NC: North Carolina Teacher Academy, 2003.

Brunsman, Bethany A. "Review of the DIBELS: Dynamic Indicators of Basic Early Literacy Skills, Sixth Edition." In *The Sixteenth Mental Measurements Yearbook,* edited by Robert A. Spies and Barbara S. Plake. Lincoln, NE: The Buros Institute of Mental Measurements, 2005.

Buck, Julie, and Joseph K. Torgesen. "The Relationship between Performance on a Measure of Oral Reading Fluency and Performance on the Florida Comprehensive Assessment Test (Tech Rep. no. 1)." Tallahassee, FL: Florida Center for Reading Research, 2003.

Dynamic Measurement Group. "DIBELS 6th Edition Technical Adequacy Information (Tech. Rep. no. 6)." Eugene, OR: Dynamic Measurement Group, 2008.

Elliot, Jacquelyn, Steven W. Lee, and Nona Tollefson. "A Reliability and Validity Study of the Dynamic Indicators of Basic Early Literacy Skills-Modified." *School Psychology Review,* vol. 30, no. 1, 2001, pp. 33-49.

Fantuzzo, John W. and Virginia Hampton. "Penn Interactive Peer Play Scale: A Parent and Teacher Rating System for Young Children." In *Play Diagnosis and Assessment (2nd Edition),* edited by K. Gitlin-Weiner, A. Sandgrund and C. Schaefer. New York: Wiley, 2000.

Farrell, Linda, Carrie Hancock, and Susan Smartt. *DIBELS: The Practical Manual Book.* Frederick, CO: Sopris West, 2006.

Francis, David J., Kristi L. Santi, Christopher Barr, Jack M. Fletcher, Al Varisco, and Barbara R. Foorman. "Form Effects on the Estimation of Students' Oral Reading Fluency using DIBELS." *Journal of School Psychology,* vol. 46, 2008, pp. 315-342.

Good, Roland H., Deb Simmons, Edward J. Kame'enui, Ruth A. Kaminski, and Joshua Wallin. "Summary of Decision Rules for Intensive, Strategic, and Benchmark Instructional Recommendations in Kindergarten through Third Grade (Technical Report no. 11)." Eugene, OR: University of Oregon, 2002.

Good, Roland H., and Gretchen Jefferson. *Contemporary Perspectives on Curriculum-Based Measurement Validity.* New York: Guilford Press, 1998.

Good, Roland H., and Ruth A. Kaminski. *Catch them Early, Watch them Grow!! Using DIBELS in Your School.* Frederick, CO: Sopris West, 2003.

Good, Roland H., and Ruth A. Kaminski. *Dynamic Indicators of Early Literacy Skills (6th Ed.).* Eugene, OR: Institute for the Development of Educational Achievement, 2007.

Good, Roland H., and Ruth Kaminski. *Dynamic Indicators of Early Literacy Skills (6th Ed.).* Frederick, CO: Sopris West, 2003.

Good, Roland H., and Ruth Kaminski. *Dynamic Indicators of Basic Early Literacy Skills (5th Ed.).* Eugene, OR: University of Oregon, 2001.

Good, Roland H., and Ruth A. Kaminski. "Dynamic Indicators of Basic Early Literacy Skills (2000-2003)." Available at [http://dibels.uoregon.edu/]. 2002.

Good, Roland H., Ruth A. Kaminski, Mark R. Shinn, John Bratten, Deborah Laimon, Michelle Shinn, Sylvia Smith, and Natalie Flindt. "Technical Adequacy of DIBELS: Results of the Early Childhood Research Institute on Measuring Growth and Development. (Technical Report no. 7)." Eugene, OR: University of Oregon, 2004.

Good, Roland H., Ruth A. Kaminski, Sylvia Smith, and John Bratten. "Technical Adequacy of Second Grade DIBELS Oral Reading Fluency Passages (Technical Report no. 8)." Eugene, OR: University of Oregon, 2001.

Good, Roland H., Josh Wallin, Deborah C. Simmons, Edward J. Kame'enui, and Ruth A. Kaminski. "System-Wide Percentile Ranks for DIBELS Benchmark Assessment (Technical Report 9)." Eugene, OR: University of Oregon, 2002.

Hagan-Burke, Shanna, Mack D. Burke, and Clay Crowder. "The Convergent Validity of the Dynamic Indicator of Basic Early Literacy Skills and the Test of Word Reading Efficiency for the Beginning of First Grade." *Assessment for Effective Intervention,* vol. 31, no. 4, 2006, pp. 1-15.

Hintze, John M., Amanda L. Ryan, and Gary Stoner. "Concurrent Validity and Diagnostic Accuracy of the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of Phonological Processing." *School Psychology Review,* vol. 32, no. 4, 2003, pp. 541-556.

Kame'enui, Edward J. "An Analysis of Reading Assessment Instruments for K-3." University of Oregon: Institute for the Development of Educational Achievement, 2002.

Kaminski, Ruth A., Roland H. Good, Mark Shinn, Sylvia R. Smith, Deborah Laimon, and Michelle Shinn. "DIBELS Word use Fluency Measure for Kindergarten through Third Grades (Technical Report no. 13)." Eugene, OR: University of Oregon, 2004.

Kaminski, Ruth A., and Roland H. Good. "Toward a Technology for Assessing Basic Early Literacy Skills." *School Psychology Review,* vol. 25, 1996, pp. 215-227.

National Reading Panel. *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and its Implications for Reading Instruction.* Washington, DC: National Institute of Child Health and Human Development, 2000.

Powell-Smith, Kelly A., Roxanne Hudson, Jose M. Castillo, and Robert Dedrick. "Examining the use of DIBELS Nonsense Word Fluency with First and Second Grade Students in Reading First Schools." Manuscript submitted for publication, 2008.

Riedel, Brant W. "The Relation between DIBELS, Reading Comprehension, and Vocabulary in Urban First-Grade Students." *Reading Research Quarterly,* vol. 42, no. 4, 2007, pp. 546-567.

Ritchey, Kristen D. "Assessing Letter Sound Knowledge: A Comparison of Letter Sound Fluency and Nonsense Word Fluency." *Exceptional Children,* vol. 74, no. 4, 2008, pp. 487-506.

Roberts, Greg, Roland H. Good, and Stephanie Corcoran. "Story Retell: A Fluency-Based Indicator of Reading Comprehension." *School Psychology Quarterly,* vol. 20, no. 3, 2005, pp. 304-317.

Roehrig, Alysia D., Yaacov Petscher, Stephen M. Nettles, Roxanne F. Hudson, and Joseph K. Torgesen. "Accuracy of the DIBELS Oral Reading Fluency Measure for Predicting Third Grade Reading Comprehension Outcomes." *Journal of School Psychology,* vol. 46, 2008, pp. 343-366.

Rouse, Heather L., and John W. Fantuzzo. "Validity of the Dynamic Indicators for Basic Early Literacy Skills as an Indicator of Early Literacy for Urban Kindergarten Children." *School Psychology Review,* vol. 35, no. 3, 2006, pp. 341-355.

Shanahan, Timothy. "Review of the DIBELS: Dynamic Indicators of Basic Early Literacy Skills." In *The Sixteenth Mental Measurements Yearbook,* edited by Robert A. Spies and Barbara S. Plake. Lincoln, NE: The Buros Institute of Mental Measurements, 2005.

Shapiro, Edward S., Emily Solari, and Yaacov Petscher. "Use of a Measure of Reading Comprehension to Enhance Prediction on the State High Stakes Assessment." *Learning and Individual Differences,* vol. 18, 2008, pp. 316-328.

Shaw, Rose, and Dale Shaw. "DIBELS Oral Reading Fluency-Based Indicators of Third Grade Reading Skills for Colorado State Assessment Program (CSAP) (Technical Report)." Eugene, OR: University of Oregon, 2002.

Schilling, Stephen G., Joanne F. Carlisle, Sarah E. Scott, and Ji Zeng. "Are Fluency Measures Accurate Predictors of Reading Achievement?" *The Elementary School Journal,* vol. 107, no. 5, 2007, pp. 429-448.

Tindal, Gerald, Doug Marston, and Stanley L. Deno. "The Reliability of Direct and Repeated Measurement (Research Rep. no. 109)." Minneapolis: University of Minnesota Institute for Research on Learning Disabilities, 1983.

Uribe-Zarain, Ximena. "Relationship between Performance on DIBELS Oral Reading Fluency and Performance on the Reading DSTP Year 2004-2006." Newark, DE: Delaware Education Research and Development Center, University of Delaware, 2007.

U.S. Department of Education. "Proven Methods: Early Reading First and Reading First." Available at [http://www.ed.gov/nclb/methods/reading/readingfirst.html]. 2006.

Vander Meer, Carolyn D., F. Edward Lentz, and Stephanie Stollar. "The Relationship Between Oral Reading Fluency and Ohio Proficiency Testing in Reading (Technical Report)." Eugene, OR: University of Oregon, 2005.

Wilson, John. "The Relationship of Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Oral Reading Fluency to Performance on Arizona Instrument to Measure Standards (AIMS)." Tempe, AZ: Tempe Arizona School District No. 3, 2005.

Wood, David E. "Modeling the Relationship Between Oral Reading Fluency and Performance on a Statewide Reading Test." *Educational Assessment,* vol. 11, no. 2, 2006, pp. 85-104.

EARLY CHILDHOOD LONGITUDINAL STUDY–KINDERGARTEN CLASS OF 1998–1999 (ECLS–K) MATHEMATICS ASSESSMENT, 2004

Authors: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education		**Type of Assessment:** Individually administered adaptive assessment (requires computer administration) **Domain:** Mathematics
Publisher: National Center for Education Statistics (NCES) http://nces.ed.gov/ECLS/ Contact first for general permission to use and then contact publisher for use of certain copyrighted items:[1] Riverside Publishing 800-323-9540 http://www.riverpub.com Pearson 800-627-7271 http://www.pearsonassessments.com PRO-ED, Inc. 800-897-3202 http://www.proedinc.com		**Grade/Age Range:** Kindergarten through grade 5 **Administration Interval:** None described but conducted fall and spring
Material, Training, and Scoring Costs: Easels, computer administration programs, and training materials were developed for the study; researchers should coordinate with NCES about availability of these materials as resources. Development costs for assessments vary with number of students assessed and waves of data. Publisher costs for use of copyrighted items must be negotiated ($1 or less per student).		**Personnel and Training Requirements** **Credentials Required for Use:** Level 3 (licensure or state certification, doctorate) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (> 2 hours)
Languages: English, Spanish		**Alternate Forms:** No
Representativeness of Norming Sample: The base sample (those with a parent interview or child assessment in either fall or spring) was drawn from over 1,000 schools within 100 primary sampling units (PSU) across the United States to be nationally representative of students in kindergarten in 1998–1999. The scores are based on a sample of 11,200 students in the ECLS–K grade 5 round (90 percent in grade 5).[2] Testing was conducted in 2004 but with collection of longitudinal data on the ECLS–K mathematics assessment starting in 1998.		**Summary** **Initial Material Cost:** Not available **Time to Administer:** 30 minutes **Ease of Administration and Scoring:** 4 (administered or scored by clinician or specialist) **Reliability:** 3[3] (all at or above 0.70) **Predictive Validity:** Available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 3 (normed within past 10 years and nationally representative)[1]

NARRATIVE

Description: The Early Childhood Longitudinal Study–Kindergarten Class of 1998–1999 (ECLS–K) mathematics assessment is an individually administered adaptive test of elementary school mathematics. It takes about 30 minutes and features a two-stage format, with a total of 153 items across kindergarten through grade 5.[4] Students first complete a common routing test (17 to 18 items) and, based on a predetermined cutoff score on the routing test, receive a second-stage skill-level test—low, middle, or high (about 20 to 30 items with some overlap across skill levels). The assessor uses computer-assisted personal interviewing to ask questions and enter responses. The computer is programmed for routing to the necessary second-stage form and includes standardized instructions and assessor prompts. Most items involve easel administration; in later waves, students complete some items with paper-and-pencil workbooks. The mathematics assessment covers a wide range of topics that increase in difficulty with age while measuring conceptual and procedural knowledge as well as problem-solving skills across five content strands (number, measurement, geometry, data analysis, algebra).

Other Languages: The ECLS-K developed a Spanish version of the mathematics assessment for Spanish-speaking students in kindergarten and grade 1 only. First the English mathematics assessment was translated into Spanish, which native Spanish speakers then back translated into English. Expert mathematicians, who were native Spanish speakers, then reviewed the Spanish mathematics assessment. Approximately 1,000 students completed the assessment in the fall of kindergarten, with decreasing numbers assessed in Spanish over the next two years as students were routed to the English version based on an English language proficiency screener. Developers conducted several analyses to establish the comparability of the scores between the English and Spanish versions, including differential item functioning (DIF) analysis (see Bias Analysis), comparison of actual versus predicted item performance, analysis of item response theory (IRT) model fit statistics, and an examination of gains across versions. The analyses detected individual items (under 15 out of 64) that demonstrated cultural-linguistic bias or differences between actual and predicted item performance across the language versions. Across all analyses, authors summarized the differences detected to be small and unlikely to affect ability estimates, with most items performing similarly between the English and Spanish versions (Rock and Pollack 2002).

Uses of Information: The ECLS–K mathematics assessment assesses the status and growth of mathematical knowledge and skills typically taught to students as part of elementary school curricula.

Methods of Scoring: Assessors enter student responses into a computer. The ECLS–K study provides raw scores for the number of correct responses on the specific routing test and on proficiency levels (such as three of four correct items). Scoring requires the services of a psychometrician to create a data file with the item responses coded as correct, incorrect, omitted, or not administered. The psychometrician then estimates scale scores. The ECLS–K used a three-parameter IRT model for scoring. The psychometric manuals provide the item parameters (see Pollack et al. 2005 for details on the IRT model and procedures and its Appendix B for item parameters). IRT-based criterion-referenced and standardized scores include the theta estimates (standardized scores of true ability; mean = 0, standard deviation = 1), IRT-based scale scores in the metric of the number of test items (a non-linear transformation of the theta estimates,

summing the probability of responding correctly to all 153 items for a given theta estimate), and probabilities for proficiency levels. Scores are available for nine proficiency levels across grades: (1) number and shape, (2) relative size (to include more advanced items on number), (3) ordinality/sequence (includes more advanced items on number and solving word problems), (4) addition/subtraction, (5) multiplication/division, (6) place value, (7) rate and measurement, (8) fractions, and (9) area and volume. In addition, a norm-referenced *T*-score is available (transformation of the theta; standardized mean = 50, standard deviation = 10) for performance relative to the population of students at a particular time point. The ECLS–K norm-referenced scores are available for kindergarten and grades 1, 3, and 5.

Interpretability: The ECLS–K User's Manuals provide descriptions of the raw and IRT-based scores and of the appropriate use of the various scores depending on research questions (overall versus specific skills, mastery versus relative status). A few researchers have decided to use the theta ability estimates as opposed to the IRT-based scale scores to model growth (see Hong and Yu 2007; Reardon 2007).

Reliability: Information comes from the ECLS–K study conducted between 1998 and 2004. Reliability estimates are based on approximately 18,000 students during kindergarten, a subsample of about 5,000 students in fall grade 1, 16,600 students in spring grade 1, 14,400 students in grade 3, and 11,200 students in grade 5.

(1) Internal consistency reliability: Reliability of theta (based on combined first- and second-stage tests) ranged from 0.89 to 0.94 across grades. Cronbach's alphas for raw scores, with items varying in each grade, ranged from 0.78 to 0.88 for the routing test across kindergarten and grades 1, 3, and 5; 0.66 to 0.78 for the low form; 0.58 to 0.72 for the middle-skill form; and 0.73 to 0.83 for the high-skill form. Split-half reliability estimates computed for the nine proficiency levels (of four items each) generally fell between 0.41 (level 1, fall kindergarten) and 0.68 (level 7, grade 5). Exceptions included the first two levels of spring grade 1 scores, with lower reliability estimates (0.26 to 0.32) likely due to limited variance, and one split-half reliability estimate exceeding 0.70—level 6 scores at grade 5 with a coefficient of 0.78.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: The ECLS–K study investigated the potential that individual assessor variation in administration affects student performance and introduces variance between students with the fall kindergarten and spring grade 1 data. Developers conducted a three-level hierarchical linear model (HLM) (students within assessors within study teams) to determine the proportion of variance in students' scores attributable to variance between assessors. The majority of variance in students' mathematics scores was between students (92 percent), and only about 1.5 percent was attributable to assessors, indicating little effect of assessors' administration and computer entry on differences between students' scores. In addition, during the spring grade 3 and 5 data collections, a quality control observer scored a subset of items alongside the assessor. Percent agreement ranged from 98 to 100 percent across skill level tests (one to four items each) and averaged 99 percent across all items across grades.

Validity Evidence:
The development of the ECLS–K mathematics assessment involved an ongoing process that included background review of instruments (for example, state assessments), national and state

performance standards, and the National Assessment for Educational Progress (NAEP) framework as well as the expert recommendations of curriculum experts, teachers, and academics on grade-appropriate and developmentally important content. Researchers field-tested item pools at various times with students from different grade levels. Classical item analyses included percent correct (to assess item difficulty), corrected item-total score correlations (to assess item discrimination), distracter analysis (for suitability of response options for selected response items), and the number of students refusing to respond to the item (to suggest confusion). Developers conducted IRT scaling for item selection purposes as well as for estimating scale scores (see Methods of Scoring). Developers noted a review items for demonstrating satisfactory psychometrics across these various analyses, eliminating items of concern unless deemed necessary for framework specification. In addition to the item analyses conducted as part of the field test for the final selection of items, the ECLS–K psychometric reports reviewed the findings from the final ECLS–K data collection as compared to other studies (for example, NAEP) in areas such as achievement gaps between subgroups.

Construct/Concurrent validity: Investigation for the validation of scores of the ECLS–K mathematics assessment involved correlating it with the Woodcock-McGrew-Werder Mini-Battery of Achievement (MBA) Calculation and Reasoning and Concepts subtests. Correlations between the MBA total score and the ECLS–K mathematics theta score ranged from 0.80 to 0.84 for students in grade 2 through 5 as part of the ECLS–K field tests (approximately 300 students in each grade). The main ECLS–K study also presented a correlation of 0.65 at grade 5 between the theta score and the teacher's academic rating of mathematics skills (approximately 5,000 students).[5]

Correlations between the ECLS–K mathematics and reading assessments ranged from 0.74 to 0.77 across kindergarten through grade 5. Correlations between the ECLS–K mathematics and science assessments were 0.73 and 0.75 at grade 3 and 5, respectively. Correlations at grade 5 between the ECLS–K mathematics assessment and teacher ratings of reading and science competencies were 0.58 and 0.55, respectively (approximately 10,000 and 5,000 students).[3]

The analysis showed that the IRT-predicted proportion of correct items increased over time, indicating that the ECLS–K mathematics assessment measures growth across years of schooling.

Predictive validity: Correlations of the theta score estimates between the elementary school years for students with all six rounds of ECLS–K data (fall and spring kindergarten, fall and spring grade 1, and spring grades 3 and 5) ranged from 0.70 to 0.88, with lower correlations as time between rounds increased. In addition, fall kindergarten mathematics scores predicted grade 3 achievement as measured by the mathematics assessment and teacher ratings (Duncan et al. 2007). (The interested reader may also review the ECLS web site for a study bibliography at http://nces.ed.gov/ecls/pdf/bibliography.pdf).

Bias Analysis: Developers screened the pool of items for sensitivity toward subgroups (unspecified) before pilot testing in 1999. Later, they conducted differential item functioning (DIF) analyses with field test data as part of the development of the final assessment. If a reviewer determined that differences were not tied to relevant content, the analyses led to the elimination of items from the final assessment. The number of dropped items was not consistently noted (five at grade 3; three at grade 5).

In determining the comparability of the English and Spanish mathematics assessments during kindergarten and grade 1, DIF analyses revealed differences on two to four items at different time points, favoring students taking the English mathematics assessment. Expert panels reviewed the items and determined that the content was relevant and the items appropriate for inclusion but provided no further information. The developers concluded across all analyses that the number and magnitude of differences were small enough as to not compromise the comparability of ability estimates between the two language versions of the assessment (see Other Languages).

In addition, developers conducted DIF analyses on the final assessment form in each ECLS–K data collection round; however, with IRT scores recalibrated with each round, only the DIF analyses from grade 5 are reported to date. They conducted DIF analyses for subgroups by gender and race (White versus Black, Hispanic, Asian-Pacific Islander, and other races), detecting DIF between racial/ethnic groups. However, developers reviewed items, retaining them as relevant to the construct. In particular, DIF analyses demonstrated differences between White and Black students, with two items favoring the former and two the latter group, and between White and Asian-Pacific Islander students, with two items favoring the former and one the latter group.

Training Support: Assessors need to undergo extensive training for test administration. Tourangeau et al. (2005) provide details on the procedures employed in the ECLS–K study. Researchers may contact NCES about the availability of training materials. In brief, assessors are trained over the course of several days on the full assessment battery, with approximately one day devoted to mathematics content using role play, lecture, and practice. Certification involves coding responses given during an interactive lecture, practicing administration in pairs while observed by trainers, taking a written test on scoring and administration procedures, and administering the assessment with an actual student as a trainer uses an evaluation rubric for rapport, administration, and coding of open-ended items.

Adaptations/Special Instructions for Individuals with Disabilities: The ECLS–K User's Manuals note that accommodations for students with special needs include special settings (e.g., lighting, quiet room, adaptive chair), scheduling at particular times, modifying timing to be shorter or longer (to include split sessions), presence of a health care aide, and/or use of an assistive device (e.g., brace, cane, hearing aid, voice synthesizer). However, Braille, large print, and sign language administrations are not available. To determine the need for accommodations, the ECLS–K asked the teacher a series of questions (see NCES 2001, pp. 5–10).

Alternate Forms: None.

Previous Version: None.

NCEE or REL Study Use:[6] Evaluation of Early Elementary Math Curricula

[1] Some publishers will require researchers to indicate the exact items and exact instrument in order to obtain permission for use. Researchers would need to determine the specific items for each instrument by contacting NCES.

[2] Scores were recalibrated with each round of data collection; thus, scores from earlier rounds are based on larger samples of students, but norms are older (e.g., 18,600 to 19,600 kindergarten students during 1998–1999; 16,600 grade 1 students during spring 2000; and 14,400 grade 3 students during 2002). In addition, a new ECLS–K study is under development with a cohort of kindergartners in 2010; updated norms will be available.

[3] This rating refers to the reliability for the total test scores (IRT thetas); the proficiency levels for particular skill areas encompassed some ratings below the 0.70 level.

[4] The ECLS–K study included separate assessment batteries for kindergarten through grade 1 (17 routing items, 18 to 31 items on skill tests), grade 3 (17 routing items, 24 or 25 items on skill tests), and grade 5 (18 to 19 items each on the routing and three skill tests). Developers linked batteries across grade levels by using common items (40 common items in more than one battery) as well as a grade 2 bridge study to create a common vertical scale across kindergarten through grade 5. The assessment features an adaptive format with a wide ability range of items to limit the occurrence of floor and ceiling effects. The percentage of students with perfect or near-perfect scores, or scores below chance, was generally under one percent, reducing concerns for floor or ceiling effects.

[5] Sample size is based on completion rates reported in Tourangeau et al. 2005.

[6] See Table F.1 for web address.

References:

Duncan, Greg J., Chantelle J. Dowsett, Amy Claessens, Katherine Magnuson, Aletha C. Huston, Pamela Klebanov, Linda S. Pagani, Leon Feinstein, Mimi Engel, Jeanne Brooks-Gunn, Holly Sexton, Kathryn Duckworth, and Crista Japel. "School Readiness and Later Achievement." *Developmental Psychology*, vol. 43, no. 6, 2007, pp. 1428-1446.

Hong, Guanglei, and Bing Yu. "Early-Grade Retention and Children's Reading and Math Learning in Elementary Years." *Educational Evaluation and Policy Analysis*, vol. 29, no. 4, 2007, pp. 239-261.

National Center for Education Statistics. "ECLS–K Base Year Public-Use Data Files and Electronic Codebook: User's Manual." (NCES 2001-029rev). Washington, DC: National Center for Education Statistics, U.S. Department of Education 2001.

National Center for Education Statistics. "Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS–K), Combined User's Manual for the ECLS–K Fifth Grade Data

Files and Electronic Codebooks." (NCES 2006-032). Washington, DC: National Center for Education Statistics, U.S. Department of Education, 2006.

Pollack, Judith M., Sally Atkins-Burnett, Michelle Najarian, and Donald A. Rock. "Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS–K), Psychometric Report for the Fifth Grade." (NCES 2006-036rev). Washington, DC: National Center for Education Statistics, U.S. Department of Education, 2005.

Pollack, Judith M., Sally Atkins-Burnett, Donald A. Rock, and Michael J. Weiss. "Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS–K), Psychometric Report for the Third Grade." (NCES 2005-062). Washington, DC: National Center for Education Statistics, U.S. Department of Education, 2005.

Reardon, Sean. "Thirteen Ways of Looking at the Black-White Test Score Gap." Working paper. Stanford, CA: Stanford University. Available at [http://steinhardt.nyu.edu/scmsAdmin/uploads/001/766/24%20reardon%20black-white%20gap%20march%202007.pdf]. 2007

Rock, Donald A., and Judith M. Pollack. "Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS–K), Psychometric Report for Kindergarten through First Grade." (NCES 2002-05). Washington, DC: U.S. Department of Education, National Center for Education Statistics, 2002.

Tourangeau, Karen, Mike Brick, Lauren Byrne, Thanh Lê, Christine Nord, Jerry West, and Elvira Germino Hausken. "Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS–K), Third Grade Methodology Report." (NCES 2005-018). Washington, DC: U.S. Department of Education, National Center for Education Statistics, 2005.

Tourangeau, Karen, Thanh Lê, and Christine Nord. "Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS–K), Fifth-Grade Methodology Report." (NCES 2006-037). Washington, DC: U.S. Department of Education, National Center for Education Statistics, 2005.

EXPRESSIVE ONE-WORD PICTURE VOCABULARY TEST, THIRD EDITION (EOWPVT), 2000

Author: Rick Brownell, editor		**Type of Assessment:** Individually administered adaptive assessment **Domain:** Language arts/language proficiency (expressive oral language skills, vocabulary)
Publisher: Academic Therapy Publications 800-422-7249 http://www.academictherapy.com		**Grade/Age Range:** 2 years through 18 years, 11 months **Administration Interval:** None described
Material, Training, and Scoring Costs: The EOWPVT English Edition Test Kit includes a manual, test plates, and 25 record forms, in a vinyl folder: $155 The Spanish-Bilingual Edition Test Kit includes an EOWPVT-SBE manual, test plates, 25 Spanish-bilingual record forms, in a portfolio: $149 Scoring software: $25 when purchased with Test Kit, $45 otherwise (software may be used with both the English and Spanish-Bilingual editions)		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2 (bachelor's degree with coursework in measurement and domain) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (> 2 hours) Assessors need to be trained under the supervision of a professional familiar with educational and psychological assessment and interpretation in a relevant field (e.g., psychology). In addition, the assessor should administer several practice trials.
Languages: English, Spanish		**Alternate Forms:** No
Representativeness of Norming Sample: The norms were based on a nationally representative school-age sample of 2,327 individuals age 2 through 18 years, 11 months. It included 60 to 228 children per age group in 12-month intervals for age 2 to 14 years (for example, 60 2-year-olds, 105 3-year-olds, and 209 4-year-olds), 124 15- to 16-year-olds, and 119 17- to 18-year-olds selected to match the population distribution of the 1990 U.S. Census. The sample was stratified by age, region, race/ethnicity, parent education, community size, gender, and disability status. Norming sample participants were included only if English was their primary language. Testing was conducted in 1999 in 32 states.[1]		**Summary** **Initial Material Cost:** 2 ($100 to $200) **Time to Administer:** 10 to 15 minutes **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 3 (normed within past 10 years and nationally representative)

NARRATIVE

Description: The EOWPVT is an individually administered adaptive test that measures English expressive vocabulary—normed for age 2 through 18 years, 11 months. It includes test plates or pictures that the student must identify orally. The measure has 170 items, and the average administration time is 10 to 15 minutes while scoring averages less than 5 minutes. Items on the test are ordered by increasing difficulty, with a basal of eight consecutive correct responses and a ceiling of six consecutive incorrect responses.

Other Languages: The EOWPVT-Spanish Bilingual Edition (SBE) is normed on a national sample of Spanish-bilingual individuals, age 4 through 12 years, 11 months. Fifty percent of the sample had mothers with less than a high school diploma (Brownell 2001). Basal and ceiling rules also differ from the English edition. Record forms for the Spanish-Bilingual edition include acceptable responses in both English and Spanish. Both assessments have the same number of items and are considered comparable measures. For additional information on the EOWPVT-SBE, please contact the publisher.

Uses of Information: The EOWPVT is used to measure an individual's vocabulary, identify difficulties in reading or expressing words, screen pre-kindergarten and kindergarten students based on vocabulary skills, and assess the English vocabulary of an English language learner. The assessment may also be used to measure expressive aphasia by testing students on both the EOWPVT and the Receptive One-Word Vocabulary Test (ROWPVT). In addition, the developers state that the measure may be used to evaluate cognitive ability.

Methods of Scoring: Assessors score items on a pass/fail basis depending on the verbal response given by the student (more than one answer may be acceptable) and obtain the raw score by adding the number of correct responses (all responses below the basal are counted as correct). Raw scores may be converted into age-adjusted standard scores, percentile ranks, age equivalents, Normal Curve Equivalents, *T*-scores, scaled scores, and stanines. Tables for obtaining the scores listed above are available in the manual, although the scoring software offers an alternate method.

Interpretability: An individual with formal training in psychometrics should interpret EOWPVT scores. The scoring software may be used with both the English and Spanish-Bilingual editions of the EOWPVT, and a score difference analysis is available for students given both assessments. The scoring software also provides a summary of converted scores and a graph of the results based on word responses or raw scores.

Reliability: Analyses on reliability are based on a standardization study including 3,661 individuals (Berry et al. 2004). The correlations below are uncorrected, though the manual also provides correlations corrected for restricted range or variance.
(1) Internal consistency reliability: Cronbach's alpha coefficients ranged from 0.93 to 0.98, and split-half reliability coefficients ranged from 0.92 to 0.97 across one-year age intervals between 2 and 14 years and two-year intervals between 15 and 18 years.
(2) Test-retest reliability: The correlations between the scores of two administrations (with an average of 20 days between tests) ranged from 0.85 to 0.97 across ages (0.85 for age 2 through 4

years, 0.92 4 through 7 years and 8 through 10 years, 0.87 11 through 13 years, and 0.97 14 through 18 years). A sample of 226 students was retested.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: Reliability of scoring, response evaluation, and administration were used to evaluate inter-rater reliability.
Reliability of scoring compared hand-scored tests of four assessors (two with experience and training on the EOWPVT and two with no prior experience or training but given the scoring instructions from the manual). Each scored the same 30 randomly selected examinations (2 for each of the 15 age groups). The scores were then compared to software scoring of the assessments, with agreement across all scorers at 100 percent.
With *reliability of response evaluation*, correct or incorrect indications were omitted from copies of the 30 examinations mentioned above, and a trained assessor scored the examinations by using the written responses, which were then compared to the original scores. Out of 2,508 responses, scores obtained by the trained assessor and the 30 original assessors matched 99.4 percent.
Reliability of administration examined the consistency of test administration across assessors. During a single testing session, students were administered the assessment twice by two assessors. A single assessor then scored both administrations. The authors found a correlation of 0.95 between the two scores across administrations. The students sampled (N = 20) ranged from age 3 through 17 years.

Validity Evidence:
The third edition of the EOWPVT combines previous editions—the lower level (age 2 to 11 years) and the upper extension (age 11 to 15 years)—and extends its use through 18 years. Development of the EOWVPT for previous editions surveyed parents for common words used by young children and then compiled common words used in the home, community, and school for inclusion on the assessment. The third version maintains many of these items. The addition of new items was based on the use of dictionaries and other vocabulary resources.

In October and November 1998, a pilot test was conducted to establish item difficulty and item validity of the three levels of the assessment—the lower level, upper extension, and the new items—in order to combine each into a single form for the third edition of the EOWPVT. The original item order was maintained for the lower level and upper extension, with the new items ordered according to perceived difficulty. The rank order correlation of the initial/final order was 0.95. A sample of 154 students age 2 through 18 years was used.

Classical Test Theory (CTT) and Item Response Theory (IRT) were used during the item selection process. Item analyses resulted in the elimination of four items (additional items were eliminated per the Bias Analyses discussed below). The correlation of item difficulty to item order using the 170 items on the final assessment was 0.99. Correlations across age groups ranged from 0.93 to 0.98, with a median of 0.96. The discrimination index was 0.91, with a range of 0.75 to 0.99 across age groups and a median of 0.81. The entire standardization sample was used.

Construct/Concurrent validity: Data from the standardization study, including 3,661 individuals, were used for all validity analyses (Berry et al. 2004). The correlations below are uncorrected, though the manual provides correlations corrected for restricted range or variance.

The EOWPVT has been compared to other vocabulary assessments and to assessments measuring language and academic achievement. Among expressive and receptive vocabulary measures such as the Peabody Picture Vocabulary Test (PPVT), Expressive Vocabulary Test (EVT), and Receptive One-Word Vocabulary Test (ROWPVT), correlations between scores ranged from 0.62 to 0.83. The five language assessments (e.g., Oral and Written Language Scales) measured expressive and receptive language, listening and auditory comprehension, and grammar. Correlations with these assessments ranged from 0.36 to 0.81 for subtest scores and from 0.43 to 0.76 for total scores. Four achievement assessments (e.g., Metropolitan Achievement Test) measuring reading and language correlated from 0.41 to 0.71 for reading scores and from 0.46 to 0.63 for language scores. In addition, correlations with vocabulary subtest scores of two intelligence tests—the Wechsler Intelligence Scale for Children-Third Edition and Stanford-Binet Intelligence Scale-Fourth Edition—ranged from 0.64 to 0.78. Sample sizes were below 70, except for comparison to the ROWPVT, which used the same norming sample as the EOWPVT; student age ranged from 2 through 18 years. Correlations between scores from previous editions of the EOWPVT and the current version were 0.79 and 0.82 for 1979 and 1990, respectively.

EOWPVT scores are distinguishable by age, ability, and disability status. The correlation between age and the EOWPVT raw score was 0.84 for age 2 through 18 years, consistent with the assumption that the extent of an individual's expressive vocabulary increases with age. For a sample of 40 students age 7 through 17 years, scores on the Otis-Lennon School Ability Test (OLSAT-7), which measures abstract thinking and reasoning, correlated with the EOWPVT scores at 0.74 and 0.39 for the OLSAT verbal and nonverbal scores, respectively. To compare students by disability status, *t*-tests were used to compare standard scores of 1,023 individuals identified with one or more types of disabilities to the estimated population mean standard score of 100. The first 8 of the 11 disability groups (mental retardation, autism, language delay, expressive/recessive language disorder, behavioral disorder, learning disabilities, hearing loss, auditory processing deficit, ADHD/ADD, articulation, fluency disorder) were significantly lower than the mean.

Predictive validity: No information available.

Bias Analysis: An analysis of differential item functioning (DIF) was conducted by using the Mantel-Haenzel procedure for the following subgroups: gender (male versus female), residence (urban versus rural), and race/ethnicity (White students versus each Black, Hispanic, and, Other racial/ethnic students). To select the items for this third edition, item analysis and differential item functioning statistics were considered in addition to suggestions from assessors and members of a cultural review panel. An additional 12 items were eliminated based on these considerations. A sample of 2,945 students age 2 through 18 years was used.

Training Support: Assessors without coursework in measurement should be trained and supervised by a professional familiar with educational and psychological assessment and interpretation. The manual provides thorough instructions for administration and scoring of the assessment. The assessor should administer several practice trials before administering the assessment.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: None.

Previous Version: The third edition of the EOWPVT combines previous editions developed for age 2 through 11 years (lower level) and age 11 through 15 years (upper extension) and extends the use of the assessment through age 18. In addition, it reflects new national norms and was co-normed with the Receptive One-Word Picture Vocabulary Test (ROWPVT). With the third version, items were added and dropped (i.e., biased or outdated items), illustrations were modified for clarity and updated to full color, and administration procedures were revised to include assessor prompts and cues.

NCEE or REL Study Use:[2] National Evaluation of Early Reading First

[1] The EOWPVT and Receptive One-word Picture Vocabulary Test (ROWPVT) were co-normed.

[2] See Table F.1 for web address.

References:

Berry, Daniel J., Lisa J. Bridges, and Martha J. Zaslow. "Early Childhood Measures Profiles." Washington, DC: Child Trends, 2004.

Brownell, Rick (ed.). *Expressive One-Word Picture Vocabulary Test: Spanish-Bilingual Edition.* Novato, CA: Academic Therapy Publications, 2001.

Brownell, Rick (ed.). *Expressive One-Word Picture Vocabulary Test Manual—Third edition*. Novato, CA: Academic Therapy Publications, 2000.

Brownell, Rick. *Expressive One-Word Picture Vocabulary Test-Scoring Software.* Novato, CA: Academic Therapy Publications, 2000.

Longo, Alfred. "Review of EOWPVT." In *The Fifteenth Mental Measurements Yearbook,* edited by Barbara S. Plake, James C. Impara, and Robert A. Spies. Lincoln, NE: The Buros Institute of Mental Measurements, 2003.

EXPRESSIVE VOCABULARY TEST,
SECOND EDITION (EVT-2), 2007

Authors: Kathleen T. Williams		**Type of Assessment:** Individually administered adaptive assessment **Domain:** Language arts/language proficiency (vocabulary)
Publisher: Pearson Assessments (800) 627-7271 http://ags.pearsonassessments.com/		**Grade/Age Range:** 2 years, 6 months to 90 years **Administration Interval:** None described
Material, Training, and Scoring Costs: EVT-2 Form A or B Test Kit (includes manual, 25 record forms for each form, easel, and carrying case): $215 each or $390 for both EVT-2 ASSIST scoring CD-ROM: $259		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2 (bachelor's degree with coursework in measurement and domain) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (> 2 hours) Assessors should be trained and experienced in test administration and interpretation and should have practiced administering the EVT-2.
Languages: English		**Alternate Forms:** Two forms; no administration interval described
Representativeness of Norming Sample: The sample consisted of a stratified random sample of 3,540 individuals ages 2.5 to over 90 years (between 100 and 200 each at one-year intervals for ages 2 to 22 years) selected to match the U.S. population proportionately on gender, race/ethnicity, parents' average education level, geographic region, and special education status. The sample was restricted to individuals proficient in English. A subsample of 2,003 students in kindergarten through grade 12, based on U.S. Census data on grade distribution, was used to establish grade norms. The assessments were conducted from fall 2005 to spring 2006 at 320 sites nationwide.		**Summary** **Initial Material Cost:** 3 ($200 to $500) **Time to Administer:** 10 to 15 minutes **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 3 (normed within past 10 years and nationally representative)

NARRATIVE

Description: The Expressive Vocabulary Test, Second Edition (EVT-2), an individually administered adaptive assessment to determine oral vocabulary and word retrieval skills for standard English, is appropriate for ages 2 years, 6 months to 90 years and above. With the latest update, the EVT-2 now has two parallel forms, each with 190 items ordered by increasing difficulty and age-appropriate practice items. During the 10- to 15-minute assessment, the assessor presents a color picture on an easel and reads aloud a stimulus question, allowing about 10 seconds for a response. The assessor administers the items beginning at a predetermined age-appropriate start item until the basal and ceiling items are found. The basal item is the lowest of five consecutive correct items (going to preceding items in reverse order if necessary); the ceiling item is the highest of five consecutive incorrect items.

Other Languages: None.

Uses of Information: The EVT-2 is used to assess vocabulary acquisition (status and growth) by measuring expressive vocabulary and word retrieval skills. Authors note that the EVT-2 may also be used for (1) screening for expressive-language problems, (2) screening preschool children's development, (3) measuring word retrieval when used with the Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4; see Interpretability), (4) exploring reasons for reading difficulties, and (5) evaluating development of English vocabulary in non-native speakers (though it cannot provide a normative score for such individuals).

Methods of Scoring: Assessors receive a list of all acceptable correct answers for open-ended questions and a list of the most common incorrect responses. The raw score is calculated as the total number of incorrect items subtracted from the number of the ceiling item. Raw scores may then be converted into normative scores and growth scale value (GSV) scores by using a series of tables. Normative scores (using norms by age; grade: fall; or grade: spring) include standard scores and associated confidence intervals, percentiles, normal curve equivalents, stanines, and age and grade equivalents. Alternatively, computer software (EVT-2 ASSIST) is available to score and interpret assessment results.

Interpretability: Brief statements explain the interpretation of the possible derived scores, and the authors encourage consideration of other factors, including health, other assessment results, and direct observations during the assessment, in interpreting the results. Authors note that assessors should interpret scores for students with special characteristics only if so qualified. Authors also note that a raw score of 0 cannot be accurately standardized and interpreted, and raw scores of 190 (the highest possible) should be interpreted with caution. The EVT-2 age norm score may be used in combination with age norm scores from the PPVT-4 to help determine whether the EVT score is indicative of word retrieval problems. The manual provides possible explanations for significantly different scores between the two measures as well as other possible qualitative interpretations of categories of items on the EVT-2, such as home versus school vocabulary or by part of speech.

Reliability:
(1) Internal consistency reliability: The Spearman-Brown split-half reliability (within forms) ranged from 0.88 to 0.95 for Form A scores and from 0.89 to 0.95 for Form B scores for those

age 2 years, 6 months to 24 years. Cronbach's alphas for the same age groups ranged from 0.94 to 0.98 for Form A scores and from 0.93 to 0.97 for Form B scores. Calculation of split-half reliabilities was based on separate analysis of the odd and even items using a Rasch analysis. The correlations between forms were adjusted for differences in the standard deviations of the normative sample for each form.
(2) Test-retest reliability: The correlation coefficients between scores from the two administrations ranged from 0.94 to 0.96. The interval between the two administrations ranged from 2 to 8 weeks for students age 2 to 14 years (N = 348). No information was provided for individuals between ages 15 and 22 years.
(3) Alternate form reliability: The correlation coefficients between scores from the two forms administered within one session or in two sessions up to 7 days apart ranged from 0.79 to 0.91 for students age 2 to 14 years (N = 507). No information was provided for individuals between ages 15 and 22 years.
(4) Inter-rater reliability: No information available.

Validity Evidence:
The EVT-2, as a measure of oral or expressive vocabulary knowledge and word retrieval, does not require reading, writing, or lengthy responses. It includes words only if they have high or moderately high frequency and are not learned only through specialized training. Frequency was determined by using several frequency word lists, including *The Reading Teacher's Book of Lists, Fourth Edition* (2000), *The American Heritage Word Frequency Book* (1971), and *A Spoken Word Count* (1966), among others. Stimulus words were grouped into 20 descriptive categories and evaluated during two national try-outs. The manual details decisions guiding word selection and picture development for stimulus words, construction of the second parallel form, and scoring criteria. Classical and Rasch item analyses were used to gauge item difficulty. The manual details how words were determined to be correct or incorrect responses. The authors state that the detailed content specifications and item development process provides qualitative evidence of the content validity of the EVT-2 as a measure of standard American English expressive vocabulary.

Construct/Concurrent validity: Studies correlated the EVT-2 scores with scores from four instruments that measure vocabulary, language ability, and/or reading achievement: the PPVT-4; the Comprehensive Assessment of Spoken Language (CASL); the Clinical Evaluation of Language Fundamentals, Fourth Edition (CELF-4); and the Group Reading Assessment and Diagnostic Evaluation (GRADE). In addition, scores on the EVT were correlated with those on the EVT-2. Typically, instruments were administered on the same day, except for the EVT, which was given up to 11 days later. Samples generally included 100 to 450 students, except for the PPVT-4, which used the same norming sample of 3,540 individuals. Students ranged in age from 2 to 24 years but typically were in elementary or middle school. Correlations between the EVT-2 and PPVT-4 scores ranged across grades from 0.80 to 0.84. Correlation coefficients with CASL subtest scores ranged from 0.45 to 0.83. Correlations with the CELF-4 language subtest scores ranged from 0.67 to 0.80. Correlations with the GRADE ranged across grades from 0.51 to 0.74 on the total scores and from 0.35 to 0.73 on vocabulary and comprehension composite scores. Correlations with scores on the previous edition of the EVT ranged across grades from 0.76 to 0.83.

All tests of difference of means between 12 student groups were statistically significant (controlling for gender, race/ethnicity, and socioeconomic status). Groups included giftedness, special education disabilities, or language delay groups and a non-clinical reference group from the norming sample.

Predictive validity: No information available.

Bias Analysis: During pre-release development trials, the developers conducted item bias analysis with respect to gender, race/ethnicity, socioeconomic status, and regional location, using a sample of 1,451 individuals age 2 years, 6 months to 21 years, with Black and Hispanic students overrepresented. The developers also paid attention to students with special education status or other disabilities or delays. During the first national try-outs, the developers eliminated or revised and re-tested any items that exhibited potential bias and dropped from the test re-tested items whose performance in the second national trial did not improve. Items also underwent review for cultural sensitivity with respect to fairness and appropriateness.

Training Support: Pearson Assessments offers in-service training and content presentations, some in person and some online.

Adaptations/Special Instructions for Individuals with Disabilities: Specific modifications are presented for individuals with hearing problems depending on their method of communication. Any adaptations made for impairments should be noted and considered in interpretation, as should any atypical student behavior. The authors caution that such problems and adaptations may invalidate the normative score obtained from the assessment because individuals with similar impairments were not part of the norming sample.

Alternate Forms: The EVT-2 has two parallel forms—Form A and Form B. Administration intervals between forms were not described.

Previous Version: The developers made several changes from the EVT to the EVT-2. They created a parallel form with new content; modernized vocabulary and included words learned in home life and everyday living skills; added the early literacy vocabulary that is used to instruct young students; added GSV score computations; expanded the initial set of labeling exercises and now include the exercises throughout the measures; printed stimulus questions on the form for each item to facilitate administration; and dropped less representative or dated items.

NCEE or REL Study Use:[1] The Effectiveness of a Program to Accelerate Vocabulary Development in Kindergarten (REL-Southeast)

[1] See Table F.1 for web address.

References:

Pearson Assessments. "EVT-2 Publication Summary Form." Available at [http://ags.pearsonassessments.com/pdf/pubsum/evt2.pdf]. 2007.

Montgomery, Judy K. *The Bridge of Vocabulary: Evidence-Based Activities for Academic Success.* Minneapolis, MN: Pearson Assessments, 2007.

Pearson Assessments. *ASSIST: Automated System for Scoring and Interpreting Standardized Tests*. Minneapolis, MN: 2006.

Williams, Kathleen T. *Expressive Vocabulary Test—Second Edition Manual*. Minneapolis, MN: NCS Pearson, Inc., 2007.

GATES-MACGINITIE READING TESTS, FOURTH EDITION (GMRT-4), 2002

Authors: Walter H. MacGinitie, Ruth K. MacGinitie, Katherine Maria, Lois G. Dreyer, and Kay E. Hughes

Type of Assessment: Group-administered assessment
Domain: Reading (phonological awareness, letter sounds, reading vocabulary, reading fluency, comprehension skills)

Publisher:
Riverside Publishing Company
800-323-9540
http://www.riverpub.com

Grade/Age Range: Kindergarten through adult
Administration Interval: Fall and spring; more frequently (i.e., winter) if alternate forms used

Material, Training, and Scoring Costs:
Online versions available (Level 1 through AR): $4 per student initial administration, $1 per student retest
Level PR through 3: Test package for hand scoring (25): $88 (both forms); for machine scoring (25): $146 Level PR; $123 Level BR through 3 each
Level 4 through AR: Reusable test booklet packets: $88 (both forms)
Hand-scoring answer sheets (25): $45.79
Machine-scorable answer sheets (100): $125
Scoring options: Software ($475; Johnson 2005), by publisher for additional cost, or hand-scored with acetate scoring template ($28.39 each level)

Personnel and Training Requirements
Credentials Required for Use: Level 2 (bachelor's degree with coursework in measurement and domain)
Personnel for Administration: Individual with basic clerical skills with some training
Training for Administration: Minimal (1 to 2 hours)

Languages: English

Alternate Forms: Two forms (Level 2 through AR only); possible to conduct several tests through the school year

Representativeness of Norming Sample:
The GMRT-4 was normed through standardization studies conducted in 1998 and 1999 with about 65,000 kindergarten through grade 12 students and 2,800 adults from across the United States. The sample was nationally representative (stratification based on variables obtained from Quality Education Data, which included geographic region, district enrollment, and socioeconomic status), but racial/ethnic and gender breakdowns are not described. Renorming research was conducted to provide updated 2006 norms for each level, but details on the standardization process and sample are not yet publicly available.

Summary
Initial Material Cost: 2 ($100 to $200)
Time to Administer: 55 to 100 minutes depending on level
Ease of Administration and Scoring: 3 (administered and scored by a highly trained individual)
Reliability: 3 (all at or above 0.70)
Predictive Validity: Available
Construct/Concurrent Validity: Available[1] **Norming Sample Characteristics:** 3 (normed within past 10 years and nationally representative)

NARRATIVE

Description: The Gates-MacGinitie Reading Tests, Fourth Edition (GMRT-4) is a group-administered assessment of students' reading skills for kindergarten students through adults. It consists of 11 levels: Pre-Reading (PR) and Beginning Reading (BR) for kindergarten, separate assessments for grades 1 through 6, Level 7/9, Level 10/12, and Adult Reading (AR), each with two to four subtests. The assessment takes 55 minutes to administer for Levels 1 and 3 through AR and 75 minutes at Level 2 because of an additional subtest. The assessment also takes longer to administer at the PR and BR levels (75 to 100 minutes) as the assessor must read the questions aloud to the students; for the higher levels, students read and respond to questions silently, following the instructions read by the assessor (detailed in the Directions for Administration manual included in each test package). All students mark their answers in a booklet. GMRT-4 is a paper-and-pencil assessment but is also available for online administration.

Level PR assesses beginning reading skills and the early conceptual development behind them. It consists of 90 items across four subtests: (1) Literacy Concepts on understanding of words and phrases commonly used in beginning reading instruction; (2) Oral Language Concepts (Phonological Awareness), focusing on phonemic units; (3) Letters and Letter/Sound Correspondences; and (4) Listening (Story) Comprehension assessing students' ability to understand connected text. Level PR is for students who are about to learn to read in kindergarten or grade 1 (grade range listed as K.7 to 1.4); therefore, the answer choices for Level PR are mostly pictures. Students mark their answer choices directly in a test booklet.

Level BR assesses students' decoding skills and consists of 70 items across four subtests: (1) Initial Consonants and Consonant Clusters; (2) Final Consonants and Consonant Clusters; (3) Vowels; and (4) Basic Story Words (identifying commonly used English words that do not require decoding). Response options consist of both pictures and words. Level BR is designed for students at the beginning and end of grade 1 (grade range listed as 1.0 to 1.9).

Levels 1 and 2 assess students' independent reading skills. Both levels include a Word Decoding subtest and a Comprehension subtest (using extended written text). The grade range for Level 1 is 1.5 to 1.9. Level 2 also contains a Word Knowledge subtest that assesses beginning reading vocabulary. Although students taking Levels 1 and 2 may be in the same testing room, Level 2 consists of 125 items (compared to 82 in Level 1) and an additional subtest.

Levels 3 through 10/12 assess reading achievement in grades 3 through 12. Each level consists of 93 items with a Vocabulary subtest and a Comprehension subtest. For the vocabulary subtest, students must select the option word or phrase with the closest meaning to the tested word. Students taking Level 4 through 10/12 may be in the same testing room.

Level AR is for use in post–high school education programs to assess the reading achievement of students enrolled in such programs. The format is the same as for Level 3 through 10/12.

All test levels are designed for the given grade or grade range. According to the Directions for Administration, however, levels are usually considered suitable for students at the beginning of the following grade (if they are average or below) and students at the end of the previous grade (if they are above average).

Other Languages: None.

Uses of Information: The GMRT-4 assesses students' reading skills from kindergarten through college and describes the reading achievement of individuals and groups of students. The developers suggest that specific types of norm-referenced scores (such as normal curve equivalents [NCE], stanines, and percentile ranks) may be used to identify students who are advanced or delayed in reading and therefore need additional attention.

Methods of Scoring: Each level is supported by a Manual for Scoring and Interpretation that provides information on scoring procedures and the types of available scores. A variety of scoring options is available—hand scoring, machine scoring with software, or machine scoring by the publisher (see Material costs for various booklet types or http://www.riverpub.com/products/gmrt/scoring.html). Raw total and subtest scores are calculated by summing the correct responses and may be converted into five types of norm-referenced scores: NCEs, percentile ranks, stanines, grade equivalents (GE), or extended scale scores (ESS) (Johnson 2005). For each test level, the Manual for Scoring and Interpretation provides norming tables to convert raw scores to norm-referenced scores. Renorming research provided updated 2006 norms for each level, but details on the standardization process are not yet publicly available. A document is provided to convert 1999 norms to 2006 norms (MacGinitie et al. 2007).

Interpretability: The developers provide a Linking Teaching to Testing manual and a Manual for Scoring and Interpretation to help users interpret the raw and derived scores. The developers do not specify qualifications for interpretation.

Reliability:
(1) Internal consistency reliability: The developers provide Kuder-Richardson Formula 20 (KR-20) coefficients for raw total and subtest scores across grade levels and forms for two time points. For Levels PR and BR, coefficients for the total scores ranged from 0.93 to 0.95 and, for subtest scores, from 0.79 to 0.89. For Levels 1 and 2, coefficients for the total scores ranged from 0.96 to 0.97 and, for subtest scores, from 0.92 to 0.94. For Level 3 through 10/12, coefficients for the total scores ranged from 0.93 to 0.96 and, for subtest scores, from 0.90 to 0.93.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: The developers conducted an equating study in 1999 for parallel forms. They calculated reliability coefficients between Forms S and T by using a sample of over 10,000 students for Level 2 through AR (grades 2 through 12 and college). Based on a sample of over 1,100 students, coefficients for total and subtest scores on Level 2 were 0.95 (total), 0.92 (Word Decoding), 0.90 (Word Knowledge), and 0.86 (Comprehension). Coefficients for total scores on Level 3 through AR ranged from 0.81 to 0.93, using student samples ranging from 67 to over 1,400 students across grades. Coefficients for subtest scores on Level 3 through AR ranged from 0.75 to 0.90 (Vocabulary) and from 0.74 to 0.89 (Comprehension).
(4) Inter-rater reliability: No information available.

Validity Evidence:
The developers conducted field testing, bias review and analysis, and an extensive process of question selection to develop and norm the GMRT. The process involved estimating the

difficulty level of each question (item p-values) and item discrimination indices (biserial correlations). For those levels (2 through AR) for which alternate forms were available, the developers closely matched items across forms on item difficulty, item discrimination, and Harris-Jacobson grade-level ratings to ensure equivalence. The developers used the resultant information along with input from teachers, former teachers, and field-testing results to ensure appropriate content and balance of questions at each level. The developers also conducted three equating studies to ensure overlap between adjacent test levels, equivalence between Forms S and T, and equivalence between the third and fourth editions of the GMRT. Over 30,000 students from across the United States participated in the three equating studies.

Construct/Concurrent validity: The GMRT-4 Technical Report refers the reader to the GMRT-3 manual for information on validity studies comparing the GMRT-3 with the PSAT Verbal, SAT Verbal, ACT English, and other reading assessments.[1]

Authors cite the equating study between the third and fourth editions of the GMRT as additional evidence of validity. Correlations between GMRT-4's Level PR and BR total scores and GMRT-3's Level PRE and R were each 0.89, with samples of 1,032 and 423 students, respectively. Correlations between the two editions' Level 1 through 10/12 ranged from 0.82 to 0.93 for total scores, from 0.77 to 0.90 for Vocabulary subtests, and from 0.58 to 0.88 for Comprehension subtests. Sample sizes ranged from 43 to over 1,200 students across grades.

Predictive validity: The GMRT-4 Technical Report describes correlations calculated between scores from fall and spring administrations of Form S for Level PR through 10/12 with students in the standardization sample. For Level PR through 1 (grade 1), correlations ranged from 0.66 to 0.90 for total raw scores based on samples ranging from 78 to 610 students across levels (subtest correlations not provided). For Level 2 (grade 2), correlations between administrations were 0.90 for total scores and ranged from 0.82 to 0.86 for subtests (N = 906 students). For Level 3 through 10/12, correlations ranged from 0.71 to 0.93 for total raw scores, from 0.75 to 0.91 for Vocabulary subtests, and from 0.58 to 0.86 for Comprehension subtests. Sample sizes ranged from 87 to 601 students across levels. The developers also provide means for comparison across time points.

Bias Analysis: To detect questions that could be biased, the developers conducted differential item functioning (DIF) analysis, comparing Black and Hispanic students to all other students by using the Mantel-Haenszel procedure and then eliminating questions with a "strong suggestion of DIF." Before deciding whether to retain or eliminate a question, content experts reviewed questions that met certain statistical criteria of effect size and statistical significance, thus indicating DIF. In addition, a diverse set of consultants (such as sociolinguists) from a variety of ethnic and professional backgrounds conducted a bias review to identify any questions that could be considered offensive or biased.

Training Support: The Directions for Administration provide assessors with detailed instructions on how to use practice items with students, how to deal with problems during testing, where to start and stop, how to handle make-up testing for absentees, and how to prepare for administration (e.g., scheduling, environment). The developers state that assessors should read the Directions for Administration and the instructions before beginning administration of the assessment.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: The GMRT-4 consists of two alternate forms: Forms S and T. Form S is available for all levels (PR through AR), and Form T is available for Level 2 through AR. The suggested administration interval for each form is fall and spring, but additional administrations (i.e., winter) are possible when alternate forms are used.

Previous Version: The developers made several changes to the GMRT-4, including the addition of subtests, the renaming of some subtests, and the addition of new test levels. They added new subtests to Levels PR and BR for Listening (Story) Comprehension and Basic Story Words, respectively, and a new subtest to Level 2 for Word Knowledge. They also revised the Comprehension subtests in Levels 1 and 2 so that three or four consecutive questions make up a short story, and they split the third edition's Level 5/6 test into two levels. The developers added a new test level for adult reading (Level AR), with community college norms and expanded "out-of-level" norms for testing in grades above or below the recommended grade level for each level of the test.

NCEE or REL Study Use:[2] Impact of the Thinking Reader Software Program on Grade 6 Reading Comprehension, Vocabulary, Strategies, and Motivation; Evaluation of Principles-Based Professional Development to Improve Reading Comprehension for English Language Learners (REL-Pacific); Assessing the Impact of Collaborative Strategic Reading (CSR) on Reading Comprehension

[1] The GMRT-4 Technical Report refers the reader to the GMRT-3 manual for information on validity studies. Attempts to obtain the third edition report were unsuccessful.

[2] See Table F.1 for web address.

References:

Johnson, Kathleen M. "Review of Gates-MacGinitie Reading Tests, Fourth Edition Forms S and T." In *The Sixteenth Mental Measurements Yearbook*, edited by Robert A. Spies and Barbara S. Plake. Lincoln, NE: The Buros Institute of Mental Measurements, 2005.

MacGinitie, Walter H., Ruth K. MacGinitie, Katherine Maria, and Lois G. Dreyer. *Gates-MacGinitie Reading Tests Fourth Edition: Directions for Administration, Level 2, Forms S and T*. Rolling Meadows, IL: Riverside Publishing, 2000.

MacGinitie, Walter H., Ruth K. MacGinitie, Katherine Maria, and Lois G. Dreyer. *Gates-MacGinitie Reading Tests Fourth Edition: Technical Report Forms S and T*. Itasca, IL: Riverside Publishing, 2002.

MacGinitie, Walter H., Ruth K. MacGinitie, Katherine Maria, Lois G. Dreyer, and Kay E. Hughes. *Gates-MacGinitie Reading Tests Fourth Edition: Manual for Scoring and Interpretation, Level 2, Forms S and T*. Rolling Meadows, IL: Riverside Publishing, 2007.

MacGinitie, Walter H., Ruth K. MacGinitie, Katherine Maria, Lois G. Dreyer, and Kay E. Hughes. *Gates-MacGinitie Reading Tests Fourth Edition: Score Comparisons 1999 to 2006 Norms*. Rolling Meadows, IL: Riverside Publishing, 2007.

McCabe, Patrick P. "Review of Gates-MacGinitie Reading Tests, Fourth Edition Forms S and T." In *The Sixteenth Mental Measurements Yearbook*, edited by Robert A. Spies and Barbara S. Plake. Lincoln, NE: The Buros Institute of Mental Measurements, 2005.

GROUP READING ASSESSMENT AND DIAGNOSTIC EVALUATION (GRADE), 2001

Author: Kathleen T. Williams		**Type of Assessment:** Group-administered assessment, individual assessment **Domain:** Reading (reading vocabulary, comprehension)
Publisher: Pearson Education, Inc. 800-321-3106 http://www.pearsonschool.com		**Grade/Age Range:** Preschool through postsecondary **Administration Interval:** Two to three months
Material, Training, and Scoring Costs: GRADE Classroom Sets with Form A (materials to assess 30 students, including Student Booklets, Teacher's Administration Manual, and Teacher's Scoring & Interpretive Manual; for Levels 4–6, also includes Hand-Scoring Templates and Answer Sheets): $127.50 to $200.95, depending on level Grade Classroom Sets with Forms A and B: $220.50 to $323.50, depending on level Technical Manual: $35.50 Out-of-Levels Norms Supplement: $35.50 Scoring and Reporting Software (required for out-of-level testing): $411.95 (Hand Entry Version) or $2,341.95 (Scanning Version)		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2 (bachelor's degree with coursework in measurement and domain) **Personnel for Administration:** Individual with basic clerical skills with some training **Training for Administration:** Minimal (1 to 2 hours)
Languages: English		**Alternate Forms:** Yes, two parallel forms; two to three months administration interval
Representativeness of Norming Sample: Standardization data were collected from two nationwide samples in 2000. The first sample consisted of 16,408 preschool through grade 12 students at 122 sites. The second sample consisted of 17,024 preschool through postsecondary students. The sample approximated U.S. census levels with respect to region, gender, community size and type, and socioeconomic status (although it contained fewer low-income students and more high-income students than the overall population). The race/ethnicity of students also approximated U.S. levels: 63.5 percent of the students were White, 17.5 percent were Black, 15.5 percent were Hispanic, and 3.5 percent had other backgrounds. The sample included mainstreamed students with disabilities.		**Summary** **Initial Material Cost:** 2 ($100 to $200) **Time to Administer:** 45 to 90 minutes **Ease of Administration and Scoring:** 2 (self-administered or administered and scored by someone with basic clerical skills) **Reliability:** 3[1] (all at or above 0.70) **Predictive Validity:** Available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 3 (normed within past 10 years and nationally representative)

NARRATIVE

Description: The Group Reading Assessment and Diagnostic Evaluation (GRADE) is a norm-referenced, paper-and-pencil assessment of pre-reading and reading skills. It may be administered to groups or individual students from preschool through postsecondary (young adult) levels. The GRADE consists of 11 grade-based test levels: Level P (prekindergarten and kindergarten); Level K (kindergarten and grade 1); Level 1 (kindergarten through grade 2); Levels 2 through 6 (for in-level or out-of-level testing, grades 2 through 6); Level M (grades 5 through 9); Level H (grades 9 through 12); and Level A (grades 11, 12, and postsecondary). Each level has subtests with developmentally appropriate task and skill demands. For example, Level P subtests focus on pre-reading skills and tasks, whereas Level A subtests focus on advanced vocabulary and grammar, inference, and synthesizing information. The assessment's 16 subtests are grouped into five main components: (1) Pre-Reading; (2) Reading Readiness, (3) Vocabulary; (4) Comprehension, and (5) Oral Language. The components (and subtests within components) that are assessed at each level vary according to developmentally appropriate task and skill demands. Level P consists of four Pre-Reading subtests and two Reading Readiness subtests. Level K is comprised of six Reading Readiness subtests and one Vocabulary subtest. Levels 1 through A include Sentence Comprehension, Passage Comprehension, and Listening Comprehension as core subtests; additional level-specific subtests in the Vocabulary component include Word Meaning (Levels 1 and 2), Word Reading (Levels 1, 2, and 3), and Vocabulary (Levels 3 through A). Students must attempt all items, the number of which varies by level. For the lowest five levels, students mark their answers in scannable test booklets that include test items and pictures. Students at the higher levels use reusable test booklets and mark their answers on separate answer sheets. There are two assessment forms per level. Administration times vary by test level, with older students typically completing the assessment in 45 to 60 minutes. For younger students, the author recommends short breaks during the test session, which typically expands the testing time to 60 to 90 minutes.

Other Languages: None.

Uses of Information: Researchers may use the GRADE to assess the effectiveness of teaching and intervention methods. The parallel forms may be used for pre- and post-testing in evaluations of reading remediation or enrichment programs. Researchers can also use the GRADE's growth score values to collect longitudinal data on a common metric across a wide range of grade levels. In educational settings, the GRADE may be used for planning and placement, for monitoring growth over time and diagnosing reading strengths and weaknesses relative to the norming sample, and for out-of-level testing with students exhibiting exceptional reading abilities or difficulties. Schools may use the GRADE to meet the assessment requirements of federal and state programs such as Reading First and Early Reading First.

Methods of Scoring: To obtain raw scores, the assessor may use hand-scoring templates, answer keys, or GRADE Scoring and Reporting Software (the answer keys are included in the GRADE Teacher's Scoring & Interpretive Manual for each level; the hand-scoring templates and software are sold separately). Subtest raw scores are the total number of correct answers in that subtest. The assessor records raw scores in the Score Box on the front of the answer sheet or, if answers were marked in the Student Booklets, on the Class or Individual Score Summary worksheets that may be reproduced from the manual. The Score Box includes places to record raw scores for

each subtest assessed (subtests vary across levels), composite scores (a Vocabulary Composite for Levels 1 through 3 and a Comprehension Composite for Levels 1 through A), and the total test score, which is the sum of varying subtest and composite scores for each level. The assessor converts the raw scores to normative scores by using the appropriate norms table, as determined by (1) the test level and form, (2) whether scores will be based on fall or spring norms, (3) the student's grade, and (4) whether the testing was on or out of grade level. The Teacher's Scoring & Interpretive Manual includes norms tables for each level; scoring out-of-level testing results requires the Out-of-Level Norms Supplement booklet or GRADE Scoring and Reporting Software (both sold separately). The assessor converts raw scores for all subtests, composite scores, and the Total Test score into stanines, and also derives percentiles, grade equivalents, standard scores, and normal curve equivalents (NCE) for the composite scores, total test score, and the Vocabulary subtest score (for Levels 4 through A only). Finally, the assessor records a total test growth scale value (GSV). The GSV quantifies reading achievement on an equal-interval scale and compares it to the entire range of achievement across all grades as opposed to scores that compare student achievement to that of students in a particular grade (e.g., stanines, percentiles, and standard scores). The GSV may be used as a common metric for tracking growth across grade levels.

Interpretability: The Teacher's Scoring and Interpretive Manual provides information on how to make norm-referenced interpretations for each type of score and includes instructions for summarizing GRADE scores for groups of students or individuals. It includes a reproducible Class Score Summary sheet for recording and comparing scores of up to 25 students as well as an Individual Score Summary sheet for use in communicating results to parents or other teachers. Diagnostic analyses of group and individual performance may also be recorded on reproducible worksheets for each subtest. The worksheets allow for comparison of a student's performance on individual items to that of peers in the national sample. The diagnostic information may also be used to analyze a student's performance across subtests and item types. The author cautions that extremely high or low raw scores (only one or two items missed or correct) signal that the test level was not appropriate for that student and that the results will have limited diagnostic value.

Reliability:
(1) Internal consistency reliability: Across test levels, forms, and samples (spring and fall), Cronbach's alpha coefficients for total scores ranged from 0.89 to 0.98. Corresponding split-half reliability estimates (corrected by the Spearman-Brown Prophecy formula for full-test length) ranged from 0.94 to 0.99. For subtest, composite, and total scores at each GRADE test level, alpha and split-half reliability coefficients ranged from 0.33 to 0.99; 94 percent of the coefficients were equal to or greater than 0.70. Of the coefficients below 0.70, almost all were on the optional Listening Comprehension subtest.
(2) Test-retest reliability: Researchers administered the same form of the appropriate GRADE test level twice to 816 students drawn from the fall standardization sample. Across levels (P through A), coefficients for scores between the two administrations of Form A ranged from 0.77 to 0.98; for the groups taking Form B, they ranged from 0.83 to 0.96. Seventy-five percent of the sample took Form A both times; the remainder took Form B. Mean intervals between administrations ranged from 3.5 to 42 days. Correlations were corrected for restriction of range. The Technical Manual does not report uncorrected coefficients.
(3) Alternate form reliability: Corrected for restriction of range, alternate form reliability coefficients ranged from 0.81 to 0.94 across levels and order of administration. The sample

included 696 preschool through grade 12 students living in the Northeast (7.6 percent), North Central (34.4 percent), and South (58.0 percent) regions of the United States. Males and females were equally represented, and most students were White (83.9 percent), followed by Black (7.5 percent) and Hispanic (6.3 percent). Mean intervals between administrations ranged from 8 to 32 days. The Technical Manual does not report uncorrected coefficients.
(4) Inter-rater reliability: Not information available.

Validity Evidence:
The structure and content of the GRADE reflect the consensus among reading experts that ". . . learning to read progresses by a series of stages or benchmarks" (White 2001, p. 10). These stages are sequential and overlapping. The GRADE is designed to assess five components of learning to read: (1) pre-reading (visual skills and conceptual knowledge), (2) reading readiness (phonemic awareness, letter recognition, sound-symbol matching, and print awareness), (3) recognizing and understanding print vocabulary, (4) sentence and passage comprehension, and (5) acquiring complex oral language skills. The GRADE assesses this progression of skills through 11 grade-based test levels. Growth curve data presented in the Technical Manual demonstrate the progression in pre-reading and reading skills as measured by GRADE Levels P through A.

Construct/Concurrent validity: Subsamples of students in the standardization sample completed the GRADE and group-administered, nationally standardized achievement assessments. The first study included 185 students in grades 4, 5, 7, and 8 whose total scores on the GRADE were correlated with total reading scores on the Iowa Test of Basic Skills (ITBS). Correlation coefficients, corrected due to restriction of range, ranged from 0.69 to 0.83 across test levels. A second study correlated total scores on the GRADE to total reading scores on the California Achievement Test (CAT) with a sample of 119 grade 1 and 2 students. Corrected correlations between scores were 0.82 and 0.87 for grade 1 and 2 subgroups, respectively (uncorrected correlations were not reported). Researchers also correlated GRADE total scores to total scores on the Gates-MacGinitie Reading Test, a group-administered, nationally standardized reading assessment. Both assessments were administered to 313 students in grades 1, 2, 3, and 6 drawn from the standardization sample. Corrected correlation coefficients ranged from 0.86 to 0.90 across grade-level groups. In another study, 30 grade 5 students (drawn from the standardization sample) were assessed with the GRADE Level 5 and the Reading Recognition and Reading Comprehension subtests of the Peabody Individual Achievement Test-Revised (PIAT-R). Corrected correlation coefficients between GRADE scores (Vocabulary subtest, Comprehension Composite, and total test) with PIAT-R reading subtest and total scores ranged from 0.68 to 0.80.

The author reported a correlation of 0.47 between the GRADE Comprehension Composite and the PIAT-R General Information subtest, which does not require the student to read (requires listening and oral responding), concluding support for divergent validity. For 118 grade 7 and 8 students, correlations between GRADE scores (Vocabulary subtest, Comprehension Composite, and total test) and the ITBS subtests that require reading (Reading, Language Arts, and Mathematics Concepts and Problems) ranged from 0.63 to 0.83. In contrast, correlations between the GRADE scores and ITBS Mathematics Computation subtest scores were lower, ranging from 0.53 to 0.67, suggesting divergence between the measured constructs.

The Technical Manual reports differences in GRADE scores between students with reading difficulties and matched control groups. In one study, GRADE scores of 242 dyslexic students (grouped into four GRADE test levels) were compared to those of a random sample of control students (drawn from the standardization sample and matched on GRADE test level, gender, and race/ethnicity). The sample included students in grades 1 through 8 and some postsecondary students. Mean standard scores were significantly lower for the four dyslexic groups compared to the control groups. A second study compared the GRADE scores of 191 students diagnosed with reading disabilities and a random sample of matched control students drawn from the standardization sample. The sample consisted of students in grades 2 through 12 who were grouped by GRADE test levels. Mean scores for the learning-disabled groups were significantly lower than those of the control groups.

Predictive validity: The author conducted a study in which 232 grade 2, 4, and 6 students from the standardization sample completed the GRADE in the fall and the reading subtest of the TerraNova, a nationally standardized achievement battery, in the spring. Corrected correlation coefficients between GRADE total test standard scores and Terra Nova reading subtest standard scores were 0.76, 0.77, and 0.86 for the grade 2, 4, and 6 subgroups, respectively.

Bias Analysis: The developers of the GRADE evaluated pilot-test versions of the measure with a national tryout sample that included approximately equal numbers of male, female, White, Black, and Hispanic students. In all, the tryout sample included 20,893 students at 99 sites nationwide. The developers conducted differential item functioning (DIF) to investigate potential item bias. Rasch item calibrations were obtained for reference groups (males and Whites) and focal groups (females, Blacks, and Hispanics), and item difficulties across the groups were compared in order to identify statistically any items unfair to one or more groups after controlling for skill level. In addition, a panel of 27 educators representing women and racial minority groups identified items considered potentially inappropriate or unfair. Based on the statistical analyses and expert recommendations, items identified as potentially biased were changed or deleted. In addition, stimulus pictures were designed to present a balanced depiction of races/ethnicities and genders.

Training Support: Pearson Assessments offers free web-based and teleconference training in the use of GRADE software.

Adaptations/Special Instructions for Individuals with Disabilities: According to publisher, assessors may administer out-of-level testing to students whose reading skills are suspected to be more than two grades below the grade in which they are enrolled.

Alternate Forms: The Technical Manual cites evidence that the two forms for each GRADE level (A and B) are parallel in content and difficulty. The evidence is presented in terms of the classical test model (i.e., the levels of internal consistency, standard errors of measurement, and raw score distributions have the same means and standard deviations for each pair of forms at each level), and the content and item types are similar within pairs of the forms.

The author recommends a retesting interval of two to three months if the purpose of the testing is to assess growth or the effectiveness of instruction or intervention. In these cases, students may

be retested with the same or an alternate form. Immediate retesting with an alternate form is allowable when the results of the first testing may be invalid (e.g., due to student illness).

Previous Version: None.

NCEE or REL Study Use:[2] Closing the Reading Gap; Evaluation of Reading Comprehension Programs; Improving Adolescent Literacy Across the Curriculum in High Schools (Content Literacy Continuum, CLC) (REL-Midwest); Assessing the Impact of Collaborative Strategic Reading (CSR) on Reading Comprehension; The Enhanced Reading Opportunities Study

[1] Cronbach's alpha and split-half reliability coefficients for subtest, composite, and total scores at each GRADE test level encompassed some ratings below the 0.70 level (see Reliability).

[2] See Table F.1 for web address.

References:

Fugate, Mark H. "Review of the Group Reading Assessment and Diagnostic Evaluation." In *The Fifteenth Mental Measurements Yearbook,* edited by Barbara S. Plake and James C. Impara. Lincoln, NE: The Buros Institute of Mental Measurements, 2003.

Pearson Assessments. *Group Reading Assessment and Diagnostic Evaluation (GRADE) Scoring and Reporting Software Version 3.1*. Minneapolis, MN: Pearson Assessments, 2008.

Waterman, Betsy B. "Review of the Group Reading Assessment and Diagnostic Evaluation." In *The Fifteenth Mental Measurements Yearbook*, edited by Barbara S. Plake and James C. Impara. Lincoln, NE: The Buros Institute of Mental Measurements, 2003.

Williams, Kathleen T. *Group Reading Assessment and Diagnostic Evaluation (GRADE) Teacher's Scoring & Interpretive Manual*. Circle Pines, MN: American Guidance Service, 2001.

Williams, Kathleen T. *Group Reading Assessment and Diagnostic Evaluation (GRADE) Technical Manual*. Circle Pines, MN: American Guidance Service, Inc., 2001.

IDEA ORAL LANGUAGE PROFICIENCY TEST (IPT I–ORAL ENGLISH), 2006

Authors: Wanda Ballard, Enrique Dalton, and Phyllis Tighe

Type of Assessment: Individually administered adaptive assessment
Domain: Language arts/language proficiency (oral language proficiency in English for English Language Learners [ELL])

Publisher:
Ballard & Tighe
800-321-4332
http://www.ballard-tighe.com

Grade/Age Range: Kindergarten through grade 6
Administration Interval: None described

Material, Training, and Scoring Costs:
Test Set (50 Test Booklets, Book of Test Pictures, Examiner's Manual, Technical Manual, 50 English Test Level Summaries, 50 Spanish Test Level Summaries, 10 Group Lists): $184 for each form, varying answer sheet formats available
IPT Manager 4 scoring software (optional): $269 (single-user license), $1,120 (5-user license)
In-Service Training Kit (DVD, Trainer's Program Guide, and Briefcase): $98

Personnel and Training Requirements
Credentials Required for Use: Not specified
Personnel for Administration: Highly trained individual (recommended); at a minimum, individual with basic clerical skills with some training
Training for Administration: Minimal (1 to 2 hours)

Languages: English, Spanish—see IDEA Oral Language Proficiency Test I–Spanish (IPT I–Oral Spanish) profile

Alternate Forms: Two forms; administration interval not specified

Representativeness of Norming Sample:
The assessment was renormed in 2004 with a sample of 1,551 students' ages 5 through 12 years. Most students (93 percent) were from Texas, Colorado, and North Carolina while the remainder came from California, Maryland, and Oregon. The sample included approximately twice as many kindergarten through grade 3 students as grade 4 through 6 students. Most students were Hispanic (84.5 percent), followed by Asian or Pacific Islander (6.9 percent), Black (4.7 percent), and White (3.5 percent) students. Most students were born in Mexico (51 percent) or the United States (41 percent), and 87 percent spoke Spanish as their first language. Males and females made up 52 and 48 percent of the sample, respectively. The authors do not describe students' economic backgrounds or levels of English proficiency.

Summary
Initial Material Cost: 2 ($100 to $200)
Time to Administer: 14 minutes (on average; see Description)
Ease of Administration and Scoring: 2 (self-administered or administered and scored by someone with basic clerical skills)
Reliability: 3[1] (all at or above 0.70)
Predictive Validity: Not available
Construct/Concurrent Validity: Available[1]
Norming Sample Characteristics: 2 (older than 10 years or not nationally representative)

NARRATIVE

Description: The IDEA Oral Language Proficiency Test I–English (IPT I–Oral English) is a standardized assessment of oral English language proficiency for students whose primary language is not English. The assessment permits a norm-referenced interpretation of results and is an individually administered, adaptive test used to assess proficiency in four domains of oral English: vocabulary, comprehension, grammar/syntax, and verbal expression (including phonology). It may be used with students in kindergarten through grade 6. Assessors administer the test orally by using a booklet of pictures. The test consists of 83 items organized into six difficulty levels; each item is designed to assess a skill area and a developmental level. Based on how far a student progresses through the difficulty levels, he or she is assigned one of six corresponding score levels, called IPT score levels (A, B, C, D, E, F). Based on a student's IPT score level and grade level, the assessor determines the student's level of oral language proficiency as Non-, Limited-, or Fluent-English Speaking (NES/LES/FES). The current version of the IPT I–Oral English was published in 2001 and then renormed in 2004.

Most students are tested from the beginning level of the test through their highest level of proficiency, which is determined by stopping rules for each level. The Examiner's Manual states that students in grade 3 through 6 who demonstrate "basic oral English skills" (as observed by the assessor or as indicated in school records) may begin the test at Level C. If, however, the student misses more than one of the first six items at that level, the assessor should begin the testing at Level B. If the student misses more than one of the first six items at Level B, the assessor should begin at the lowest level of the test. The stopping rules are based on the number of errors at a level and are printed on the test sheets. Administration times vary according to students' language proficiency and the length of their responses. The average testing time is 14 minutes, with administration times ranging from about 5 minutes for students with low English proficiency to 20 minutes or more for students with higher English proficiency.

Other Languages: The IDEA Oral Language Proficiency Test I–Spanish (IPT I–Oral Spanish) assesses students' oral language proficiency in Spanish. It is not a translation of the IPT I–Oral English; rather, the IPT I–Oral Spanish is designed to assess linguistic features unique to the Spanish language. The assessments have separate norms. (See the profile for IDEA Oral Language Proficiency Test I–Spanish [IPT I–Oral Spanish] in the current compendium.)

Uses of Information: Schools may use information from the IPT I–Oral English to meet federal and state mandates for the initial assessment of language proficiency for students who are not native English speakers in order to determine potential need for special instruction. They may also use the measure to determine if students who have received specialized instruction in English meet requirements for re-designation to a higher proficiency level and corresponding educational programming or for exit from specialized instruction. The information provided by the IPT I–Oral English may also be used to diagnose strengths and weaknesses in English oral language proficiency. Although primarily designed for use by school professionals in educational settings, the instrument may also be used by researchers and program evaluators to assess students' oral proficiency in English for longitudinal studies. It could also be used in studies designed to assess the effectiveness of educational placements and interventions for students whose primary language is not English.

Methods of Scoring: Assessors may record student responses in the Student Test Booklet, on a Diagnostic Answer Sheet (DAS), or on a Scannable Answer Sheet (SAS). With each option, student responses are recorded and scored as correct or incorrect as the assessment is administered. Guidelines on the answer sheets define acceptable responses. At the end of each level, the assessor tallies the number of errors. Following defined stop rules (see Description), the level at which testing stops is the student's IPT score level (ranging from A to F). Taking into account the student's IPT score level and grade level, the assessor uses a normative designation chart to determine the student's NES/LES/FES designation. While 2006 Examiner's Manual describes determination of the designation as the final step in scoring, a 2008 scoring addendum on the publisher's web site indicates additional scoring options to determine Listening and Speaking raw scores and normal curve equivalent (NCE) scores. Assessors may calculate Listening and Speaking scores only or combine them with scores on separate IPT Reading and Writing assessments to determine an overall language proficiency score. The manual and addendum provide hand scoring instructions, but scoring software (IPT Manager 4) is also available. Results may be typed or scanned from SASs into the program. The software produces the IPT score levels, raw scores, proficiency designation, and NCE scores as well as individual and group results.

Interpretability: The IPT I–Oral English yields three types of information about students' developing oral language competencies. First, a student's IPT score level (the level at which testing stops) provides a general indication of the student's competency level. The Examiner's Manual includes Test Level Summaries that list a sampling of oral language competencies displayed by students at each score level A through F. For example, Level B competencies include telling one's name and age, identifying familiar people and objects, using the present tense of the verb "to be," and using the "ing form of a verb," whereas Level F competencies include understanding and using comparatives, superlatives, and conditional verb tenses. Second, the normative NES/LES/FES proficiency designation that is based on a student's grade and IPT score level has practical significance for educators of ELLs for planning appropriate placements and programming. Third, the new procedures outlined in the scoring protocols yield Listening and Speaking raw scores and NCE scores that may be interpreted separately or as part of an overall language proficiency score if IPT Reading and Writing assessment scores are also available. NCE scores allow educators to compare an individual student's oral language proficiency scores across grade levels and time. The addendum specifies that the NCE scores are comparable only across Forms E and F of the IPT I–Oral English, not across all IPT assessments.

Reliability:
The reliability studies described in the Technical Manual were conducted with data collected in spring 2000 during field testing of Forms E and F. The sample included 891 students in kindergarten through grade 6 who resided in 12 states in various regions of the United States. Most of the students (76 percent) were Hispanic, and 67 percent of the students spoke Spanish as their primary language.
(1) Internal consistency reliability: Ballard et al. (2006c) reported a Cronbach's alpha coefficient of 0.99 for scores of Forms E and F.
(2) Test-retest reliability: The developers re-administered the same test approximately two weeks apart. The developers reported test-retest correlations of 0.85 for scores on Form E (N = 118) and 0.83 for scores on Form F (N = 129) and 0.84 between scores for all 247 students in the sample who were tested with either Form E or F (Ballard et al. 2006c).

(3) Alternate form reliability: The developers tested 306 students with alternate test forms approximately two weeks after initial testing. Of the 306 students, 220 were given Form E followed by Form F; the remaining students were given the forms in the opposite order. Ballard et al. (2006c) reported alternate form reliability coefficients of 0.89, 0.88, and 0.91, respectively, for Forms E and F combined, Form E followed by Form F, and Form F followed by Form E.
(4) Inter-rater reliability: No information available.

Validity Evidence:
The validity studies described in the Technical Manual were conducted in spring 2000 during field testing of Forms E and F. The primary sample included 891 kindergarten through grade 6 students in 12 states. Most of the students (76 percent) were Hispanic, and 67 percent of the students spoke Spanish as their primary language. With respect to English proficiency, 22 percent of the students were designated as fluent (FES), 55 percent as limited (LES), 18 percent as non-English (NES), and 5 percent as English–only speakers. A second sample consisted of 730 kindergarten students residing in 8 states. Data were collected from the second sample in fall 2000 to provide information about students entering kindergarten. Most of the students were Hispanic (56 percent) or White (20 percent) and spoke Spanish (56 percent) or English (27 percent) as their primary language. As reported in the Technical Manual, results of the kindergarten study consist of cross-tabulations of IPT score levels with other variables.

The developers of the original IPT I–Oral English (Forms A and B) reviewed the research literature on oral language acquisition in English-speaking students, students receiving bilingual education or English as a Second Language (ESL) instruction, and the field of linguistics. They identified oral language competencies necessary for elementary students' academic success in mainstream classrooms. Within the key domains of vocabulary, comprehension, grammar/syntax, and verbal expression (including phonology), they wrote pilot items reflecting sequences of skills and competencies. The items then underwent modification based on expert recommendations. Subsequent versions of the assessment were also modified in accordance with recommendations from experts and test users.

Construct/Concurrent validity: For Form E, Ballard et al. (2006c) reported a correlation of 0.73 between IPT score levels and teacher predictions of IPT score levels. Correlations between IPT score levels and teacher judgments of academic ability, English reading ability, and English writing ability were 0.27, 0.42, and 0.40, respectively. For Form F, the correlation between IPT score levels and teacher predictions of IPT score levels was 0.83. Correlations between IPT score levels and teacher judgments of academic ability, English reading ability, and English writing ability were 0.34, 0.43, and 0.43, respectively. For Forms E and F combined, the authors reported a correlation of 0.63 between IPT score levels and teacher ratings of students' English oral language ability. To assess the comparability of Forms C and D (alternate forms of the previous version of the assessment) to Forms E and F, Ballard et al. (2006c) reported that 91 students from the "E and F sample" were also assessed with one of the older forms. They combined the data from Forms C and D and the data from Forms E and F and reported a correlation of 0.87 between Forms C and D combined and Forms E and F combined.

Ballard et al. (2006c) reported that, for Form E, IPT score levels correlated 0.43 with both age and grade. For Form F, IPT score levels correlated 0.38 with age and 0.36 with grade. For both forms, IPT score levels increased with age and grade level.

Predictive validity: No information available.

Bias Analysis: Ballard et al. (2006c) noted that, during the development of the original IPT I–Oral English, experts in bilingual education, linguistics, and oral language development screened test items for potential bias. The developers then modified items based on the experts' recommendations and repeated the process in the development of subsequent forms.

Training Support: Training in use of the IPT I–Oral English is available through three sources. First, free online in-service training is available through the publisher's web site (http://www.ballard-tighe.com). Second, test users may purchase a "do-it-yourself" in-service training kit that includes a DVD, training guides, and reproducible materials. Third, the publisher offers free onsite in-service and train-the-trainer sessions conducted by educational sales consultants. Participants who complete the train-the-trainer sessions are certified to train other test users.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: Forms E and F are parallel forms that assess the same skills with different items. The authors recommend use of alternate forms between test administrations but do not specify a minimum retest interval. Studies of alternate forms reliability used a retest interval of approximately two weeks (see Reliability).

Previous Version: The previous version of the IPT I–Oral English (Forms C and D) was published in 1991. In 1999, the developers revised Forms C and D based on advances in theory and research in oral language development and test user recommendations. They revised or deleted some items, added new items, and updated many of the picture cues. The current version, with Forms E and F replacing Forms C and D, was published in 2001.

NCEE or REL Study Use:[2] Differential Effects of English language learner training and materials—On Our Way to English (OWE) and Responsive Instruction for Success (RISE); Project ELLA (English language/Literacy Acquisition)

[1] Reliability and validity investigations were conducted in 2000 with data collected during field studies of Forms E and F.

[2] See Table F.1 for web address.

References:

Amori, Beverly, and Enrique F. Dalton. *IDEA Oral Language Proficiency Test–Spanish (IPT I–Oral Spanish).* Brea, CA: Ballard & Tighe, 2006.

Ballard & Tighe Publishers. "Deriving Reading, Writing, Listening, Speaking, Comprehension, and Overall Scores from the IPT 2004 Tests." Available at [http://www.ballard-tighe.com/pdfs/Scoring_addendum_IPT_2004_Sept_06_2.pdf]. 2006.

Ballard, Wanda S., Enrique F. Dalton, and Phyllis Tighe. *IDEA Oral Language Proficiency Test I–English, 2nd Edition.* Brea, CA: Ballard & Tighe, 2006a.

Ballard, Wanda S., Enrique F. Dalton, and Phyllis Tighe. *IDEA Oral Language Proficiency Test I–English, 2nd Edition. Examiner's Manual.* Brea, CA: Ballard & Tighe, 2006b.

Ballard, Wanda S., Enrique F. Dalton, and Phyllis Tighe. *IDEA Oral Language Proficiency Test I–English, 2nd Edition. Technical Manual.* Brea, CA: Ballard & Tighe, 2006c.

Ballard & Tighe Publishers. *IPT Manager 4.* Brea, CA: Ballard & Tighe, 2008.

IDEA ORAL LANGUAGE PROFICIENCY TEST, 3RD EDITION (IPT I–ORAL SPANISH), 2004

Authors: Beverly Amori and Enrique Dalton		**Type of Assessment:** Individually administered adaptive assessment **Domain:** Language arts/language proficiency (Spanish oral language proficiency)
Publisher: Ballard & Tighe 800-321-4332 http://www.ballard-tighe.com		**Grade/Age Range:** Kindergarten through grade 6 **Administration Interval:** None described
Material, Training, and Scoring Costs: Test Set (50 Test Booklets, Book of Test Pictures, Examiner's Manual, Technical Manual, 50 Spanish Test Level Summaries, 50 English Test Level Summaries, 10 Group Lists): $184, varying answer sheet formats available IPT Manager 4 scoring software (optional): $269 (single-user license), $1,120 (5-user license) In-Service Training Kit (DVD, Trainer's Program Guide, and Briefcase): $98		**Personnel and Training Requirements** **Credentials Required for Use:** Not specified **Personnel for Administration:** Individual with basic clerical skills with some training **Training for Administration:** Minimal (1 to 2 hours) Assessors should demonstrate Spanish language proficiency.
Languages: Spanish, English—see IDEA Oral Language Proficiency Test I–English (IPT I–Oral English) profile		**Alternate Forms:** No
Representativeness of Norming Sample: The IPT I–Oral Spanish was renormed in 2004 with a sample of 567 students in six school districts. Most students (91.3 percent) were from Texas, the remaining from Iowa. The sample included 5- through 12-year-old students. With respect to grade levels, 28.9 percent of students were in kindergarten, 19.9 percent in grade 1, 22.8 percent in grade 2, 14.8 percent in grade 3, 7.4 percent in grade 4, 4.1 percent in grade 5, and 2.1 percent in grade 6. Males and females made up 49 and 51 percent of the sample, respectively. Nearly all of the students (99.6 percent) were Hispanic and spoke Spanish as their primary language. Most students were natives of Mexico (58.9 percent) or the United States (33.5 percent); the remaining students were natives of Central American countries. No information is provided about students' economic background.		**Summary** **Initial Material Cost:** 2 ($100 to $200) **Time to Administer:** 25 minutes (on average; see Description) **Ease of Administration and Scoring:** 2 (self-administered or administered and scored by someone with basic clerical skills) **Reliability:** 3[1] (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available[1] **Norming Sample Characteristics:** 2 (older than 10 years or not nationally representative)

NARRATIVE

Description: The IDEA Oral Language Proficiency Test I–Spanish, 3rd Edition (IPT I–Oral Spanish) is a standardized assessment of oral Spanish language proficiency for Hispanic students whose primary or secondary language is Spanish. This individually administered, adaptive test permits a norm-referenced interpretation of results. It assesses proficiency in six domains of Spanish oral language development:[2] (1) syntax (arrangement of adjectives, adverbs, nouns, and verbs as well as verb tenses); (2) morphological structure (use of inflectional endings, prefixes, suffixes); (3) lexical items; (4) phonological structure (sound discrimination and word pronunciation); (5) comprehension; and (6) oral production/pragmatics. It may be used with students in kindergarten through grade 6. Assessors administer the assessment orally by using a booklet of pictures. The assessment consists of 85 items organized into six difficulty levels; each item is designed to assess a skill area and a developmental level. Based on how far a student progresses through the difficulty levels, he or she is assigned one of six corresponding score levels called IPT score levels (A, B, C, D, E, F). A student's IPT score level and grade level determine his or her level of oral language proficiency as Non-, Limited-, or Fluent-Spanish Speaking (NSS/LSS/FSS).

Most students are tested from the lowest level through their highest level as determined by the stopping rules for each level. The Examiner's Manual states that students who demonstrate "basic oral Spanish skills" (as observed by the assessor or as indicated in school records) may begin the assessment at the level specified for their grade. If, however, the student misses more than one of the first six items at that level, the assessor begins the testing again at the previous level. The stopping rules are based on the number of errors in a level and are printed on the test sheets. Administration times vary according to students' language proficiency and the length of their responses. Average testing time is 25 minutes, with administration times ranging from about 5 minutes for students with little or no Spanish proficiency to 30 minutes for students with higher Spanish proficiency.

Other Languages: The IDEA Oral Language Proficiency Test I–English (IPT I–Oral English) assesses students' oral language proficiency in English. The IPT I–Oral English and IPT I–Oral Spanish share comparable formats but are designed to assess the unique linguistic features of their respective languages. The assessments have separate norms. (See profile for IDEA Oral Language Proficiency Test I–English [IPT I–Oral English] in the current compendium.)

Uses of Information: Schools may use information from the IPT I–Oral Spanish to meet federal and state mandates for the assessment of language proficiency of students whose primary or secondary language is Spanish in order to determine potential need for special instruction. They may also use the measure to determine if students who have received specialized instruction in Spanish meet requirements for re-designation to a higher proficiency level and corresponding instructional provisions or for exit from specialized instruction. The information may also be used to diagnose strengths and weaknesses in Spanish oral language proficiency. Although primarily designed for use by school professionals in educational settings, the instrument may also be used by researchers and program evaluators to assess students' improvement of oral proficiency in Spanish in longitudinal studies. It could also be used in studies designed to assess the effectiveness of educational placements and interventions for students whose primary language is Spanish.

Methods of Scoring: Assessors may record student responses in the Student Test Booklet or on a Scannable Answer Sheet (SAS). With each option, student responses are recorded and scored as correct or incorrect as the assessment is administered. Guidelines on the answer sheet define acceptable responses. At the end of each level, the assessor tallies the number of errors. Following defined stop rules (see Description), the level at which testing stops is the student's IPT score level (ranging from A to F). Taking into account the student's IPT score level and grade level, the assessor uses a normative designation chart to determine the student's NSS/LSS/FSS designation. Additionally, the Technical Manual includes a table for converting raw scores to standard scores and percentile ranks, which are not considered part of the standard scoring procedures. Optional scoring software (IPT Manager 4) is available. Results may be typed or scanned from SASs into the program. The software produces IPT score levels, raw scores, a proficiency designation, as well as reports of individual and group results.

Interpretability: The IPT I–Oral Spanish yields different types of information about students' developing oral language competencies. First, Test Level Summaries in the Examiner's Manual summarize competencies associated with the IPT score levels (A to F). For example, at Level B, students can tell his or her name or age; identify common people and objects; use the present tense of the verb "estar"; use plurals; use "el," "la," "un," and "una" correctly; and follow simple directions involving basic positions in space. At Level F, students can understand and name opposites of key words; use the imperfect tense of irregular verbs and the preterite, past, and present tenses of verbs; comprehend and predict the outcome of a story; and recall and retell the facts of a story. The summaries represent a sampling of competencies at each level and may be useful for explaining results to parents and teachers. Second, the NSS/LSS/FSS proficiency designation has practical significance for educators of students whose native language is Spanish. These categories of oral language proficiency are recognized in the field and are useful for planning appropriate placements and programming. Third, the Technical Manual includes tables that allow assessors to determine percentile ranks, but the authors advise against doing so given the fluid nature of children's oral language development (Amori and Dalton 2006c).

Reliability:
Reliability studies described in the Technical Manual were conducted with 1995 norming sample data. The sample comprised 948 Hispanic students in eight states in kindergarten through grade 6, 95.2 percent of whom spoke Spanish as their primary language.
(1) Internal consistency reliability: Amori and Dalton (2006c) reported an overall Cronbach's alpha of 0.99 for scores.
(2) Test-retest reliability: The authors re-administered the same test form to 126 students with an interval of approximately two weeks between administrations. They reported a test-retest reliability coefficient of 0.72 (Amori and Dalton 2006c).
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: Developers of the IPT I–Oral Spanish asked different assessors to administer the assessment for the second administration of the test-retest reliability study (see above). The reliability coefficient of 0.72 is therefore an indication of both inter-rater reliability and test-retest reliability.

Validity Evidence:
Validity studies described in the Technical Manual were conducted with 1995 norming sample data. The primary sample included 948 Hispanic students in grades 1 through 6 residing in eight

states, 95.2 percent of whom spoke Spanish as their primary language. With respect to Spanish proficiency, 72.2 percent of the students were designated as fluent (FSS), 24.7 percent as limited (LSS), and 2.7 percent as non-Spanish speaking (NSS). A second sample consisted of 299 Hispanic kindergarten students residing in four states, 85.5 percent of whom spoke Spanish as their primary language. Students varied by Spanish proficiency designation (39.9 percent NSS, 52.5 percent LSS, 7.6 percent FSS). Data were collected for the second sample in fall 1995 to provide information about students entering kindergarten.

In developing both the original version (1980) and the 1995 revision of the instrument, experts reviewed current theory and research on the topic of oral language acquisition and learning. A group of language specialists identified and classified skill areas and corresponding items representing Basic Interpersonal Communication Skills (BICS) or Cognitive Academic Language Proficiency (CALP). They also developed and categorized items according to Bloom's (1956) Taxonomy of Cognitive Development, along a hierarchy from Knowledge, Comprehension, Application, Analysis, Synthesis, to Evaluation. The instrument's content and format were designed to assess the developmental, incremental, systematic, symbolic, and social aspects of students' oral language performance and development.

Construct/Concurrent validity: Based on their analysis of the primary sample data, Amori and Dalton (2006c) reported a correlation of 0.72 between students' actual score levels and teacher predictions of student score levels on the IPT I–Oral Spanish. Correlations between IPT I–Oral Spanish score levels and teacher opinions of academic ability, Spanish reading ability, and Spanish writing ability were 0.28, 0.49, and 0.46, respectively. The authors explained that the 0.28 correlation between IPT score levels and teacher opinions of academic ability was expected because Spanish oral language ability is independent of general academic ability.

Kindergarten sample data yielded a correlation of 0.60 between teachers' predicted score levels and students' actual IPT I–Oral Spanish score levels. Further, a discriminant classification analysis of teacher predictions of student proficiency level (NSS/LSS/FSS) indicated that teachers' designations of 74 percent of kindergarten and grade 1 students and 78 percent of students in grades 2 through 6 corresponded with determined IPT I–Oral Spanish score levels. Relationships between IPT score levels and teacher opinions about students' oral Spanish proficiency in kindergarten to grade 1 and grades 2 through 6 were reported in terms of Cramer's V statistics of 0.50 and 0.57, respectively.

Additionally, primary sample IPT score levels correlated 0.49 and 0.52 with age and grade, respectively, supporting the developmental nature of language acquisition as measured by the assessment.

Predictive validity: Not information available.

Bias Analysis: No information available.

Training Support: Training in the IPT I–Oral Spanish is available from the publisher. Users may purchase a "do-it-yourself" in-service training kit that includes a DVD, training guides, and reproducible materials. The publisher also offers free onsite in-service and train-the-trainer sessions conducted by educational sales consultants. Participants who complete the train-the-

trainer sessions are certified to train other assessors. Finally, the publisher offers free online in-service training. Trainees complete online training for the IPT I–Oral English, followed by a condensed training module for the IPT I–Oral Spanish. This online training is designed to familiarize assessors with the procedures of administering and scoring the assessment.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: None.

Previous Version: The original version of the IPT I–Oral Spanish was published in 1980, and the second edition was published in 1996. The third edition consists of the same form as the second edition but provides norms updated in 2004.

NCEE OR REL Study Use:[3] Project ELLA (English language/Literacy Acquisition)

[1] Reliability and validity studies described in the Technical Manual were conducted with the 1995 norming sample data.

[2] Discussions of the content of the IPT I–Oral Spanish in the Technical and Examiner's manuals sometimes refer to four main skill areas assessed by the test (Vocabulary, Comprehension, Syntax, and Verbal Expression) that are nested within these six domains.

[3] See Table F.1 for web address.

References:

Amori, Beverly, and Enrique F. Dalton. *IDEA Oral Language Proficiency Test–Spanish (IPT I–Oral Spanish).* Brea, CA: Ballard & Tighe, 2006a.

Amori, Beverly, and Enrique F. Dalton. *IDEA Oral Language Proficiency Test–Spanish (IPT I–Oral Spanish), Third Edition. Examiner's Manual.* Brea, CA: Ballard & Tighe, 2006b.

Amori, Beverly, and Enrique F. Dalton. *IDEA Oral Proficiency Test–Spanish (IPT I–Oral Spanish), Third Edition. Technical Manual.* Brea, CA: Ballard & Tighe, 2006c.

Ballard & Tighe Publishers. *IPT Manager 4.* Brea, CA: Ballard & Tighe., 2008.

INDICADORES DINÁMICOS DEL ÉXITO EN LA LECTURA (IDEL) SEVENTH EDITION, 2006

Authors: Doris Baker, Roland Good, Nancy Knutson, and Jennifer Watson		**Type of Assessment:** Individual assessment **Domain:** Reading (Spanish; phonological awareness, letter recognition and naming, vocabulary, decoding, phonics, reading fluency, different comprehension skills)
Publisher: University of Oregon Center on Teaching and Learning (free downloadable materials) 888-497-4290 https://dibels.uoregon.edu Sopris West Educational Services (print materials) 800-547-6747 http://www.sopriswest.com Wireless Generation (handheld computer software) 800-823-1969 http://www.wirelessgeneration.com		**Grade/Age Range:** Kindergarten through grade 3 **Administration Interval:** Three times per school year for Benchmark Assessments; as often as desired for Progress Monitoring Assessments
Material, Training, and Scoring Costs: Free, reproducible downloads of materials available at https://dibels.uoregon.edu Print materials: Classroom sets (IDEL Administration and Scoring Guide, 25 Benchmark Assessment sheets, 6 Progress Monitoring Scoring Booklets, and Student Materials): $57.49 (separate set required for each grade) On-site training workshops: $1,750 per day Web-based training: $1,000 per day		**Personnel and Training Requirements** **Credentials Required for Use:** Level 1 (training or supervised experience with measurement) **Personnel for Administration:** Highly trained individual (must be a native Spanish speaker or someone comfortable conversing with a native Spanish speaker) **Training for Administration:** Extensive (> 2 hours) Developers recommend assessors attend IDEL training workshops.
Languages: Spanish, English—see Dynamic Indicators of Basic Early Literacy Skills (DIBELS) profile		**Alternate Forms:** For progress monitoring, 20 or more forms available for three subtests; administer as frequently as desired
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (< $100) **Time to Administer:** Approximately 10 to 15 minutes (1 to 3 minutes per subtest) **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 1[1] (none described) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available[2] **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: Indicadores Dinámicos del Éxito en la Lectura (IDEL) 7th Edition encompasses a set of brief, individually administered, standardized screening procedures and measures that assess the early literacy skills of students learning to read in Spanish. The measures may be used with students in kindergarten through grade 3 who are learning to read either exclusively in Spanish or through English Language Learner (ELL) instruction in Spanish and English. The IDEL measures reflect and assess the linguistic structure of the Spanish language (including phonology, orthography, syntax). They directly assess phonological awareness, the alphabetic principle, accuracy and fluency with connected text, vocabulary, and comprehension; the measures permit a criterion-referenced interpretation of scores. IDEL subtests include (1) Fluidez en Nombrar Letras (Letter Naming Fluency) (FNL); (2) Fluidez en la Segmentación de Fonemas (Phoneme Segmentation Fluency) (FSF); (3) Fluidez en las Palabras sin Sentido (Nonsense Word Fluency) (FPS); (4) Fluidez en la Lectura Oral (Oral Reading Fluency) (FLO); (5) Fluidez en el Relato Oral (Retell Fluency) (FRO); and (6) Fluidez en el Uso de las Palabras (Word Use Fluency) (FUP). FNL may be used with students at the beginning of kindergarten through the beginning of grade 1, and FSF may be used with students from the beginning of kindergarten through the end of grade 1. FPS may be used with students from the middle of kindergarten through the beginning of grade 2. FLO and FRO may be administered to students in the middle of grade 1 through the end of grade 3, and FUP may be administered to all students at any time in kindergarten through grade 3.

The IDEL framework includes Benchmark Assessments comprised of grade-appropriate selections of subtests (described above), with total administration time less than 15 minutes, and Progress Monitoring Assessments (for FSF, FPS, FLO) that assess student progress and intervention effectiveness between Benchmark Assessments. Educators may administer the Progress Monitoring Assessments as frequently as necessary, using up to 20 alternate assessment tasks to prevent practice effects.

Other Languages: Dynamic Indicators of Basic Early Literacy Skills (DIBELS) 6th Edition (Good et al. 2007) assesses basic early literacy skills of students learning to read in English (see DIBELS profile in the current compendium). The IDEL measures are not a translation of the DIBELS, though the measures are based on similar theory and research about how students learn to read in alphabetic languages such as English and Spanish.

Uses of Information: The IDEL measures were developed primarily for use with students who are native Spanish-speaking ELLs, but they may also be used with students learning to read in Spanish only. The measures may assess early reading skills in Spanish in order to (1) identify students with special instructional needs; (2) plan, evaluate, or modify instructional support for identified students; and (3) monitor progress and outcomes. IDEL data may also be aggregated to track and compare the progress of groups of students and to monitor the effectiveness of reading instruction and interventions in classrooms, schools, and districts.

Methods of Scoring: Assessors must hand-score IDEL assessments. The administration and scoring guide (Cummings et al. 2006) includes guidelines for scoring responses as correct or incorrect. Assessors calculate raw scores based on the number of correct responses and compare the number of correct responses with decision rules (based on cut points). Alternatively, schools

or districts that upload local sample data into the DIBELS Data System at the University of Oregon may calculate local percentile ranks and generate custom reports and summaries based on their own data.

Interpretability: IDEL benchmark goals and related cutoff scores facilitate criterion-referenced interpretation of scores. According to the authors, benchmark goals for each measure and time period indicate the probability of achieving the next benchmark goal (goals and cutoffs have not yet been established for the FUP and FRO subtests). Students whose scores on one or more IDEL measures fall at or above the benchmark have at least an 80 percent chance of meeting the next benchmark goal (an indicator of appropriate progress). Baker et al. (2007) specify categories describing students' need for support. Based on subtest scores, assessors may categorize a student's need for support as Benchmark (having met the benchmark goal and not in need of intervention), Strategic (21 to 50 percent probability of achieving the next benchmark goal and in need of additional intervention), or Intensive (20 percent or less probability of achieving the next benchmark goal and in need of substantial intervention). Schools may also examine student performance in comparison to school or district peers. The authors recommend that schools consider student performance in relation to the benchmarks rather than use percentiles; the former are predictive of future success.

Baker et al. (2007) reported floor effects for FSF at the beginning of kindergarten. Over half of sampled students scored zero on that subtest at that time.

The IDEL Data System reports results at the student, class, school, program, and district levels. Data system users may request several types of reports, including individual student profiles, class reports (name, scores, percentiles, instructional status for all students in a class), school and district summary reports (means and proficiency level across the school year for all measures), distribution reports (disaggregated results by school, class, demographics), and district norms. Users may view reports on web pages or download PDF files.

Reliability:
(1) Internal consistency reliability: No information available.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: The DIBELS/IDEL web site presents three-week alternate form reliability coefficients for scores on four subtests: 0.91 for FNL (fall of kindergarten), 0.87 for FSF (middle of grade 1), 0.76 for FPS (middle of grade 1), and 0.87 to 0.94 for FLO (middle of grades 1, 2, 3). Watson et al. (2005) reported additional alternate form reliability coefficients of 0.86 and 0.65 for scores on FNL (kindergarten) and FSF (kindergarten), respectively (the authors did not specify intervals between test sessions). Alternate form reliability information is not available for the FRO and FUP subtests. The samples that provided the reliability data included participants in the DIBELS Data System database during the 2003–2004 and 2004–2005 school years (Ns = 6,893 and 10,942 kindergarten through grade 3 students, respectively). The authors provide little information about the samples but note that the majority of students were low-performing readers.
(4) Inter-rater reliability: No information available.

Validity Evidence:[2]
The developers of the IDEL derived cut scores, decision rules, and associated instructional recommendations from longitudinal data in the DIBELS Data System (uploaded data from participating schools and districts). They collected data on 6,893 students in kindergarten through grade 3 in 39 school districts during the 2003–2004 academic year and on 10,942 students in the same grades in 170 schools in 61 districts during the 2004–2005 school year. Baker et al. (2007) stated that the majority of students in this convenience sample resided in the states of Washington, New Mexico, and Oregon and that most were assumed to be Hispanic ELL students. In describing the development of the benchmark goals, cut points, and associated instructional categories, Baker et al. (2007) acknowledged that sampling limitations compromised their efforts in empirically establishing and validating the goals, cut points, and instructional categories and their applicability to a broader population. Little information was available about the students and schools in the sample. The students' countries of origin, levels of Spanish proficiency, and the type of instruction they received (bilingual or monolingual) could all affect the meaning of the scores. The authors stated that the sample included a large number of students with "very low" Spanish skills and noted that the sample was small relative to the longitudinal predictive analyses they conducted. With these limitations, they wrote that ". . .to determine the instructional recommendations we also relied heavily on the theoretical structure and linkage of beginning reading skills to later reading outcomes in alphabetic languages, and on our experience working with Spanish-speaking students" (p. 13). In addition, cross-year longitudinal data were available for only 15 percent of the sample, and no longitudinal data were available beyond grade 3. The authors therefore based grade 3 cut scores and instructional recommendations on "theory and estimates of previous rates of progress. . ." (p. 9).

The IDEL measures assess foundational early literacy skills, including phonological awareness, the alphabetic principle, accuracy and fluency with connected text, vocabulary, and comprehension. They reflect Spanish-specific linguistic structures. For example, nonsense words in the FPS subtest reflect the frequency of syllable patterns in Spanish.

Construct/Concurrent validity: With a sample of 48 students, Watson (2004) correlated scores from the end of grade 1 from the FSF, FPS, and FLO subtests with scores from the Woodcock-Muñoz Batería-R reading subtests (Letter and Word Identification, Word Attack, Text Comprehension, Vocabulary). Correlations between scores on FSF (Phoneme Segmentation) and three of the four Batería-R subtests ranged from 0.34 to 0.51 (Letter-Word recognition scores did not correlate significantly with FSF scores). Correlations between FPS (Nonsense Word Fluency) scores and Letter and Word Identification, Word Attack, and Text Comprehension scores ranged from 0.63 to 0.72. FLO (Oral Reading Fluency) subtest scores correlated with scores on the Batería-R Letter and Word Identification, Word Attack, and Text Comprehension subtests, with correlations ranging from 0.73 to 0.80. Baker (2007) examined correlations between scores on FLO and the subtests of the Aprenda Achievement Test with a sample of 78 grade 2 students. Correlations between scores on FLO and the Aprenda Vocabulary subtest, Comprehension subtest, and total scores ranged from 0.56 to 0.64.

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: Developers strongly recommend assessors receive training provided by the Dynamic Measurement Group (DMG; http://www.dynamicmeasurement.org) or undergo training by someone who has attended training conducted by DMG. Three types of IDEL training are available (K. Petersen, personal communication, February 9, 2009): (1) on-site professional development group workshops at introductory or advanced levels; (2) four-day summer training institutes conducted in Eugene, Oregon; and (3) web-based training for individuals and groups that have completed DIBELS training and want to be trained in the use of IDEL. The introductory on-site training workshop, Entrenamiento esencial (Essential Training), provides trainees with information about the conceptual and empirical foundations of IDEL, how to administer and score the measures, and how to use IDEL information in bilingual education. The advanced on-site training workshop, Entrenamiento avanzado (Advanced Training), is for individuals who have completed the basic training and want to extend their knowledge of IDEL and learn how to train others in its use. Training sessions are conducted in English, with examples and practice in Spanish. Participants should be able to converse comfortably with native or near-native Spanish speakers.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: Alternate forms, tasks, or passages are available for FSF, FPS, and FLO for Progress Monitoring Assessments. Twenty or more alternate forms are available for each subtest at each grade level. IDEL progress monitoring materials may be downloaded for free from the DIBELS/IDEL web site. No information is available about the equivalence of IDEL forms and passages.

Previous Version: The IDEL Sixth Edition was published in 2003. The Seventh Edition includes updated items and tasks, and its benchmark cutoff scores and decision rules for instructional recommendations were developed with data collected with students participating in the IDEL Data System during the 2003–2004 and 2004–2005 academic years.

NCEE/REL Study Use:[3] Project ELLA (English language/Literacy Acquisition)

[1] The reliability rating refers to internal consistency reliability, which was required for direct assessments (see Appendix A). See the reliability section in the profile narrative for information available on alternate form reliability for selected subtests in an earlier version of the IDEL (Baker et al. 2007).

[2] Validity studies were conducted with an earlier version of the IDEL.

[3] See Table F.1 for web address.

References:

Baker, Doris Luft. *Understanding the Relation between Oral Reading Fluency and Comprehension for Students Learning to Read in Two Languages*. Eugene, OR: University of Oregon, 2007.

Baker, Doris Luft, Kelli D. Cummings, Roland H. Good, and Keith Smolkowski. "Indicadores Dinámicos del Éxito en la Lectura (IDEL: Summary of Decision Rules for Intensive, Strategic, and Benchmark Instructional Recommendations in Kindergarten through Third Grade (Technical Report No. 1)." Eugene, OR: Dynamic Measurement Group, 2007.

Baker, Doris Luft, Roland H. Good, Nancy Knutson, and Jennifer M. Watson. *Indicadores Dinámicos del Éxito en la Lectura (7a Ed.)*. Eugene, OR: Dynamic Measurement Group, 2006.

Cummings, Kelli D., Doris Luft Baker, and Roland H. Good. "Guía en Inglés para la Administración y Calificación de IDEL." In *Indicadores Dinámicos Del Éxito En La Lectura (7a Ed.)*, edited by Doris Luft Baker, Roland H. Good, Nancy Knutson, and Jennifer M. Watson. Eugene, OR: Dynamic Measurement Group, 2006.

Good, Roland H., Nancy Bank, and Jennifer M. Watson. *Indicadores Dinámicos del Éxito en la Lectura (6ta Ed.)*. Eugene, OR: Institute for the Development of Educational Achievement, 2003.

Good, Roland H., and Ruth Kaminski. *Dynamic Indicators of Early Literacy Skills (6th Ed.)*. Eugene, OR: Institute for the Development of Educational Achievement, 2007.

Watson, Jennifer. "Examining the Reliability and Validity of the Indicadores Dinámicos del Éxito en la Lectura: A Research Study." Unpublished dissertation. Eugene, OR: University of Oregon, 2004.

KAUFMAN TEST OF EDUCATIONAL ACHIEVEMENT, COMPREHENSIVE FORM, SECOND EDITION (KTEA-II), 2004

Authors: Alan S. Kaufman and Nadeen L. Kaufman		**Type of Assessment:** Individually administered adaptive assessment **Domain:** Reading (phonological awareness, letter recognition, comprehension, decoding, fluency), language arts/language proficiency (writing, spelling, oral skills), mathematics
Publisher: Pearson Assessments 800-627-7271 http://www.pearsonassessments.com		**Grade/Age Range:** 4 years, 6 months through 25 years **Administration Interval:** None described
Material, Training, and Scoring Costs: KTEA-II Comprehensive Kit (2 easels, manual, norms book, 25 record forms, 25 student response booklets, 25 error analysis booklets, 2 each of 3 Written Expression booklets, stimulus materials, administration CD, puppet, tote bag): $341.50 for one form or $613 for both forms KTEA-II Brief Kit (easel, manual, 25 record forms, and 25 response forms): $171 KTEA-II Assist Scoring: $259 KTEA-II Training Video: $128.75		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2+ (certification beyond bachelor's like a master's) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours) Assessors should be trained in test administration, scoring, and interpretation and have practiced administering the assessment before official use.
Languages: English		**Alternate Forms:** Two forms; administration interval not described
Representativeness of Norming Sample: The norming sample consisted of a random, nationally representative grade-norming sample of 2,400 students in kindergarten through grade 12 (with 140 to 220 students per grade) and an age-norming sample of 3,000 students age 4 years, 6 months through 25 years (with 80 to 220 students per age level through age 19 and 125 individuals for 20 to 22 years and 23 to 25 years). As much as possible, the age-norming sample includes students from the grade-norming sample. Assessments occurred from September 2001 through May 2003 in 39 states and the District of Columbia. The sample was stratified based on the 2001 U.S. Census Bureau's Current Population Survey for age/grade, season, gender, ethnicity, parent education, region, special education or gifted placement, and education status for individuals over age 18.		**Summary** **Initial Material Cost:** 3 ($200 to $500) **Time to Administer:** 30 to 85 minutes depending on student age (see Description) **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 3 (normed within the past 10 years and nationally representative)

NARRATIVE

Description: The KTEA-II, an individually administered adaptive assessment, measures the reading, mathematics, and written and oral language skills of students ages 4 years, 6 months through 25 years. The assessment includes an easel administration with the use of manipulatives, student response booklets, and recorded passages for some items. Average administration time varies with student age. For prekindergarten and kindergarten students, the complete assessment takes about 30 minutes. It takes 50 minutes for students in grades 1 and 2, 75 minutes for students in grades 3 through 5, and 85 minutes for students in grades 6 and above. It comprises 8 composites and 14 subtests. The following four composites (and accompanying subtests) form the Comprehensive Achievement Composite: Reading (Letter & Word Recognition and Reading Comprehension), Mathematics (Math Concepts & Applications and Math Computation), Written Language (Written Expression and Spelling), and Oral Language (Listening Comprehension and Oral Expression). Four additional composites with accompanying subtests focus on reading skills: Sound-Symbol (Phonological Awareness and Nonsense Word Decoding), Decoding (Letter & Word Recognition from Reading Composite and Nonsense Word Decoding), Oral Fluency (Associational Fluency and Naming Facility), and Reading Fluency (Word Recognition Fluency and Decoding Fluency). The 14 subtests measure all the learning disability areas outlined in the Individuals with Disabilities Education Act Amendment of 1997. The Word Recognition Fluency, Decoding Fluency, Associational Fluency, and Naming Fluency subtests are timed administrations. The applicability of the composites and subtests vary by age and grade. The assessment does not need to be administered in its entirety. Assessors may select the relevant composites or subtests to meet their assessment needs. The average administration time for the Reading Composite ranges from 10 to 20 minutes depending on student age; from 10 to 35 minutes for the Mathematics Composite; from 20 to 30 minutes for the Written Language Composite; and from 20 to 35 minutes for the Oral Language Composite.

The items within most subtests are arranged in order of increasing difficulty, and floor and ceiling rules determine the items administered to each student. The assessor begins the assessment at the level appropriate for each grade. The basal and ceiling rules vary per subtest, but, in general, the student must correctly answer a specified number of initial items to establish the basal; otherwise, the assessor changes to the level for the previous age group and administers the items until reaching the student's ceiling. The manual describes the basal and ceiling rules applicable for each subtest.

For rapid screening of individuals age 4 years, 6 months through 90 years, the KTEA-II Brief Form is available from the publisher. The average administration time ranges from 15 to 45 minutes. The KTEA-II Brief Form consists of reading, mathematics, and written expression subtests and yields norm-referenced subtest scores and a composite score. The developer notes that, as the KTEA-II Brief Form does not contain any items from the KTEA-II Comprehensive Form, it may also be used for progress monitoring with students with KTEA-II Comprehensive Form scores.

Other Languages: None.

Uses of Information: The KTEA-II is designed to assess students' achievement across several domains, identify strengths and weaknesses, measure students' progress, and evaluate the

effectiveness of interventions. The developer also states that the KTEA-II's error analysis system allows assessors to individualize instruction and inform program planning.

Methods of Scoring: The scoring methods for the subtests vary, with detailed instructions provided in the test easel and the manual. For eight of the subtests, items are scored 0 or 1 depending on whether the student answered the item correctly. The following subtests require assessor judgment when scoring: Reading Comprehension, Listening Comprehension, Written Expression, Oral Expression, and Associational Fluency. A student's responses to items on the Oral Expression and Associational Fluency subtests must be recorded verbatim because they are scored after the testing session. For the Reading and Listening Comprehension subtests, the assessor should summarize a student's response to allow for coding upon completion of the assessment if necessary. The assessor may compute age- and grade-based standard scores, age and grade equivalents, percentile ranks, normal curve equivalents, and stanines. Raw score calculations vary by subtest; the manual provides instructions and conversion tables.

Interpretability: For high-stakes decisions concerning eligibility for special education services, only persons with a background in education and psychology and well trained in test administration and statistics should interpret the results of the KTEA-II. The manual provides detailed information on how to interpret scores. In addition to scores, the assessment record form includes a section where the assessor may record student behavioral observations during administration to help in interpreting a student's results. The developers note that the KTEA-II must be interpreted in the context of the results of other assessments and background information. The error analysis system allows assessors to obtain additional information about a student's performance and areas of weakness in order to individualize instruction and inform program planning. However, the developers note that additional diagnostic testing should be conducted when error analysis identifies areas of weakness.

Reliability:
(1) Internal consistency reliability: Split-half reliabilities for scores from the Comprehensive Achievement Composite across test forms ranged from 0.95 to 0.98 for students in prekindergarten through grade 2 and from 0.97 to 0.98 for students in grades 3 through 12. The reliability coefficients for scores of students in prekindergarten through grade 2 across test forms ranged from 0.97 to 0.99 for the Reading Composite, from 0.92 to 0.96 for Mathematics, from 0.93 to 0.96 for Written Language, from 0.86 to 0.92 for Oral Language, from 0.90 to 0.96 for Sound-Symbol, from 0.96 to 0.98 for Decoding, and from 0.60 to 0.90 for Oral Fluency. The reliability coefficients for scores of students in grades 3 through 12 across test forms ranged from 0.94 to 0.97 for the Reading Composite, from 0.94 to 0.98 for Mathematics, from 0.91 to 0.96 for Written Language, from 0.80 to 0.90 for Oral Language, from 0.89 to 0.94 for Sound-Symbol, from 0.95 to 0.98 for Decoding, and from 0.81 to 0.92 for Oral Fluency.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: Both forms of the assessment were administered in alternating order to 221 students across three groups (prekindergarten through grade 1, grades 2 through 6, and grades 7 through 12) with an administration interval of 11 to 60 days (mean of 3.5 to 4 weeks). The correlations for the Comprehensive Achievement Composite ranged from 0.92 to 0.95 across age and grade and from 0.88 to 0.94 for Reading, from 0.87 to 0.93 for Mathematics, from 0.85 to 0.91 for Written Language, from 0.64 to 0.79 for Oral Language, from 0.78 to 0.89 for Sound-Symbol, from 0.91 to 0.94 for Reading Fluency, from 0.90 to 0.93 for Decoding, and

from 0.58 to 0.77 for Oral Fluency. Correlations corrected for the variability of the standardization sample are also presented.
(4) Inter-rater reliability: The developers assessed inter-rater reliability for the subtests with more subjective scoring criteria by using a sample of 50 students per grade-level group (grade 2 or 3 and grade 8) for each subtest. The grade 2 and 3 students completed Form A, and the grade 8 students completed Form B. The reliability of scoring among raters was 0.97 for Listening Comprehension across grade-level groups and ranged from 0.93 to 0.97 for Reading Comprehension, from 0.91 to 0.96 for Written Expression, from 0.82 to 0.88 for Oral Expression, and from 0.82 to 0.97 for Associational Fluency.

Validity Evidence:
Expert consultation and a literature review established content validity. The KTEA-II was piloted with three groups of students in 2000 and 2001. In the first pilot, a sample of 4,009 students took a group-administered assessment consisting of the Math Computation, Spelling, and Reading Comprehension subtests. The second pilot involved 1,002 students administered Form A or Form B of all subtests except Math Computation and Spelling. The final pilot involved 388 students administered a revised Oral Expression subtest. A panel of 34 experts provided guidance and feedback throughout the tryout phase. A joint calibration procedure standardized the two alternate forms, and linking studies were conducted on six subtests in which students were administered questions from both forms.

Construct/Concurrent validity: Item analyses conducted after the pilots led to the revision or elimination of items showing poor discrimination or differential item functioning. In addition, the scoring criteria for several subtests underwent revision following detailed analyses of item responses from the data. Using the standardization sample of students in grade 1 and above, the developers conducted a confirmatory factor analysis involving the eight subtests associated with the four main composites (Mathematics, Reading, Written Language, and Oral Language), thus confirming the validity of the theoretical factor structure.

The developers estimated intercorrelations between the KTEA-II subtest and composite scores that reflect similar traits. For prekindergarten through grade 2 students, correlations between the Reading Composite and the other composites (excluding Mathematics) ranged from 0.43 (Oral Fluency) to 0.92 (Decoding). Correlations between the Reading Composite and the other composites (excluding Mathematics) for students in grades 3 through 12 ranged from 0.41 (Oral Fluency) to 0.86 (Decoding). The developers note that the correlation between the Reading and Decoding composites is inflated owing to a shared subtest. For prekindergarten through grade 2 students, correlations between subtests within a composite (see Description for specific subtests comprising a composite) ranged from 0.81 to 0.83 for the Reading Composite, from 0.68 to 0.69 for Mathematics, from 0.40 to 0.45 for Oral Language, and from 0.51 to 0.53 for Sound-Symbol and were 0.74 and 0.85, respectively, for Written Language and Reading Fluency. Correlations between subtests within a composite for students in grades 3 through 12 ranged from 0.63 to 0.65 for the Reading Composite, from 0.65 to 0.78 for Mathematics, from 0.64 to 0.71 for Written Language, from 0.45 to 0.51 for Oral Language, from 0.46 to 0.49 for Sound-Symbol, from 0.80 to 0.81 for Reading Fluency, and from 0.32 to 0.39 for Oral Fluency.

The developers compared the KTEA-II to the following achievement assessments: the Kaufman Test of Educational Achievement, Comprehensive Form; the Wechsler Individual Achievement

Test, Second Edition; the Woodcock-Johnson Tests of Achievement, Third Edition; the Peabody Individual Achievement Test-Revised, Normative Update; and the Oral and Written Language Scales (OWLS). Except for the OWLS, the measures were administered to two groups of students (elementary and middle/high school students) with sample sizes ranging from 73 to 172 across measures and from 29 to 89 per grade level. The OWLS was administered to a sample of 53 students in kindergarten through grade 10. The correlations between the KTEA-II Comprehensive Achievement Composite and the total scores of the other achievement assessments ranged from 0.77 to 0.93 across measures and age groups. Correlations between the KTEA-II Reading Composite and the other reading subtests ranged from 0.64 to 0.90 and from 0.62 to 0.93 for the Mathematics subtests, from 0.55 to 0.86 for the Written Expression subtests, and from 0.39 to 0.75 for the Oral Language subtests. The manual presents correlations corrected for the variability of the standardization sample.

The developers also compared the KTEA-II with the following assessments of cognitive ability: the Kaufman Assessment Battery for Children, Second Edition (KABC-II); the Wechsler Intelligence Scale for Children, Third Edition (WISC-III); and the Woodcock-Johnson Tests of Cognitive Abilities (WJ III COG). The KABC-II was co-normed with the KTEA-II and administered to 2,520 students in prekindergarten through grade 12. The WISC-III and the WJ III COG were administered to 97 and 51 students, respectively, in grades 2 through 7. Correlations between the KTEA-II Comprehensive Achievement Composite and global cognitive scores of the other assessments ranged from 0.74 to 0.79. The manual presents correlations corrected for the variability of the standardization sample.

In addition, the developers compared scores between subtests measuring largely different skills. Correlations between the KTEA-II Reading and Mathematics composites for prekindergarten through grade 2 students ranged from 0.68 to 0.76 and from 0.69 to 0.71 for students in grades 3 through 12. Correlations between the KTEA-II Written Language and Mathematics composites for prekindergarten through grade 2 students ranged from 0.66 to 0.72 and from 0.67 to 0.72 for students in grades 3 through 12. Correlations between the KTEA-II Reading Composite and the Mathematics subtests or composites of the achievement assessments mentioned above ranged from 0.34 to 0.70 across measures and age groups; from 0.28 to 0.73 between the KTEA-II Mathematics Composite and the Written Expression subtests or composites; and from 0.16 to 0.66 between the KTEA-II Mathematics Composite and the Oral Language subtests or composites from the other achievement assessments. The manual presents correlations corrected for the variability of the standardization sample.

The developers examined the measure's ability to differentiate between students with special needs and a non-clinical comparison group comprising the students from the KTEA-II age-norming sample, excluding students in the special populations. The special population studies included the following categories: reading disability, mathematics disability, writing disability, mental retardation, Attention-Deficit/Hyperactivity Disorder (ADHD), emotional/behavioral disturbance, gifted/talented, and deaf or hearing impaired. Sample sizes for the students with special needs ranged from 27 (mental retardation) to 134 (reading disability) with overlap between some categories. The scores of the clinical sample on all subscales and composites were significantly lower than those of the comparison group after controlling for gender, ethnicity, and parent education, except for students with ADHD, students with emotional/behavioral disturbance, and gifted students. The scores of the students with ADHD were significantly lower

than those of the comparison group on all subtests and composites except for the Associational Fluency subtest. The scores of students with emotional/behavioral disturbance were significantly lower than those of the comparison group on all subtests and composites except for the Reading Comprehension subtest, the Oral Language Composite and related subtests, the Phonological Awareness subtest and Sound-Symbol Composite, and the Oral Fluency Composite and related subtests. The gifted sample scored significantly higher on all subscales and composites than did the comparison group.

Predictive validity: No information available.

Bias Analysis: The developers conducted differential item functioning (DIF) by using the Rasch item response theory model for the variables of gender, ethnicity, and parent education level. A small number of items demonstrating DIF were dropped from several subtests.

Training Support: A training video is available for purchase. The developers advise assessors to practice administering the assessment before official use.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: The KTEA-II has two parallel forms—Form A and Form B. A joint calibration procedure standardized the forms, and linking studies were conducted on six subtests. Administration intervals between forms were not described.

Previous Version: The KTEA-II is an update of the Kaufman Test of Education Achievement Comprehensive Form (K-TEA) published in 1985. The five existing subtests underwent revision to accommodate the KTEA-II's expanded age range. Nine new subtests were added to provide more comprehensive coverage of students' achievement; the norms were updated.

NCEE or REL Study Use:[1] Accelerating language development in kindergarten through Kindergarten PAVEd for Success

[1] See Table F.1 for web address.

References:

Bonner, Mike. "Review of the Kaufman Test of Educational Achievement-Second Edition, Comprehensive Form." In The Sixteenth Mental Measurements Yearbook, edited by Robert A. Spies and Barbara S. Plake. Lincoln, NE: The Buros Institute of Mental Measurements, 2005.

Carpenter, C. D. "Review of the Kaufman Test of Educational Achievement-Second Edition, Comprehensive Form." In The Sixteenth Mental Measurements Yearbook, edited by Robert A. Spies and Barbara S. Plake. Lincoln, NE: The Buros Institute of Mental Measurements, 2005.

Kaufman, Alan S., and Nadeen L. Kaufman. Kaufman Test of Educational Achievement-Second Edition, Brief Form. Minneapolis, MN: Pearson, 2004.

Kaufman, Alan S., and Nadeen L. Kaufman. Kaufman Test of Educational Achievement-Second Edition, Comprehensive Form Manual. Minneapolis, MN: Pearson, 2004.

Pearson Assessments, "KTEA-II Computer ASSIST™ for Comprehensive Form." Available at [http://www.pearsonassessments.com/ktea2.aspx]. 2004.

Pearson Assessments. "KTEA-II Training Video." Available at [http://www.pearsonassessments.com/ktea2.aspx]. 2004.

MACARTHUR-BATES COMMUNICATIVE DEVELOPMENT INVENTORIES (CDI), 2007

Authors: Larry Fenson, Virginia A. Marchman, Philip S. Dale, J. Steven Reznick, Donna Thal, and Elizabeth Bates		**Type of Assessment:** Parent report **Domain:** Language arts/language proficiency (expressive and receptive language skills, vocabulary, morphology)
Publisher: Paul H. Brookes Publishing Company 800-638-3775 http://www.brookespublishing.com		**Grade/Age Range:** 8 to 37 months **Administration Interval:** None described
Material, Training, and Scoring Costs: Complete Set of CDIs (Infant and Toddler forms, User's Guide and Technical Manual, package of 20 of each form): $99.95 Complete Set of CDIs and CDI III (Infant, Toddler, and 3-Year-Old forms, User's Guide): $121.95 CDI short forms available for purchase from author: $0.25 each (http://www.sci.sdsu.edu/cdi/short_e.htm)		**Personnel and Training Requirements** **Credentials Required for Use:** No special qualifications required **Personnel for Administration:** Individual with basic clerical skills and some training **Training for Administration:** Self-training (<1 hour)
Languages: English**,** Mexican Spanish		**Alternate Forms:** No
Representativeness of Norming Sample: The updated 2007 norming sample for the CDIs (Words and Gestures, Words and Sentences) included 2,550 children (8 to 30 months) in New Haven, CT; San Diego, CA; and Seattle, WA; Dallas, TX; Madison, WI; New Orleans, LA; Providence, RI; and Storrs, CT. The sample contained equivalent numbers of boys and girls, included children from a variety of racial/ethnic groups (Black, Hispanic, and Asian), and collected information about ethnicity, birth order, maternal education, and exposure to languages other than English. The CDI-III was normed separately with 356 children age 30 to 37 months from a university subject pool. The sample included a similar number of boys and girls, and maternal education levels were higher than for U.S. Census population data.		**Summary** **Initial Material Cost:** 2 ($100 to $200) **Time to Administer:** 20 to 40 minutes **Ease of Administration and Scoring:** 3 (administered and scored by highly trained individual) **Reliability:** 3[1] (all at or above 0.70) **Predictive Validity:** Available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 2 (older than 10 years or not nationally representative)

NARRATIVE

Description: The MacArthur-Bates Communicative Development Inventories (CDI) is a parent report measure of early language skills of children between the ages of 8 and 37 months. It is a paper-and-pencil questionnaire that takes about 20 to 40 minutes to complete. The full tool consists of three inventories: (1) the CDI: Words and Gestures inventory (or Infant/Level 1 form, 8 to 18 months); (2) the CDI: Words and Sentences inventory (or Toddler/Level 2 form, 16 to 30 months); and (3) the CDI-III (30 to 37 months). Each inventory consists of several subtests. Short forms are available for the CDI: Words and Gestures and the CDI: Words and Sentences Forms A and B (described further below). Parents' response options vary by subtest. For example for a vocabulary checklist, the parent chooses from response options of "understands" or "understands and says" words. In other parts, parents circle actions or gestures that their child exhibits or provide open-ended responses to questions about their child's longest utterances.

The CDI: Words and Gestures infant inventory assesses vocabulary production and comprehension and consists of two parts, each with several subtests (see Table 1 for subtest descriptions). Part I (Early Words) consists of four subtests: (1) First Signs of Understanding, (2) Phrases, (3) Starting to Talk, and (4) Vocabulary Checklist (consists of 19 semantic subcategories). Part II (Actions and Gestures) contains five subtests: (1) First Communicative Gestures, (2) Games and Routines, (3) Actions with Objects, (4) Pretending to Be a Parent, and (5) Imitating Other Adult Actions.

The CDI: Words and Sentences toddler inventory for slightly older children assesses increased vocabulary production and grammar acquisition and consists of two parts (see Table 1 for subtest descriptions). Part I (Words Children Use) contains 685 items within two subtests: (1) a Vocabulary Checklist (consists of 22 semantic subcategories) and (2) How Children Use Words. Part II (Sentences and Grammar) contains six subtests: (1) Word Endings Part I, (2) Word Forms, (3) Word Endings Part II, (4) Combining, (5) Examples, and (6) Complexity. Before the Examples and Complexity subtests are completed, parents must respond to a question about whether their child is combining words into sentences. If the response is no, the Examples or Complexity subtests need not be completed.

CDI-III is an extension of the CDIs for children age 30 through 37 months. It is a short, single-sheet tool that measures expressive vocabulary and grammar. The first component features a 100-item vocabulary checklist (including 45 words from the CDI: Words and Sentences and 55 new words). The second component consists of 13 questions about the child's word combinations (including 12 sentence pairs, of which 5 are drawn from the CDI: Words and Sentences inventory and 7 are new items). The third component consists of 12 questions, to be answered yes or no, that ask about various aspects of comprehension, semantics, and syntax.

A short version of the inventories is available. The Level 1 form (for infants) contains an 89-word vocabulary list for 8- to 16-month-olds. Two alternate versions of the Level 2 form (for toddlers) are available for 16- to 30-month-olds, both with a 100-word vocabulary checklist. The forms may be completed in about 10 minutes and are targeted for rapid assessment or parents with limited or absent literacy skills. They may be administered by an in-person parent interview (as opposed to a paper-and-pencil questionnaire), but normative data using this approach have

Table 1. CDI Subtests and Scores

Subtest (number of items)	Description	Score
CDI: Words and Gestures (Part I—Early Words)		
First Signs of Understanding (3 items)	General questions about early comprehension of familiar words and phrases	Not described
Phrases (28 items)	Comprehension of everyday phrases and routines	Phrases Understood
Starting to Talk (2 items)	Imitation and labeling	Not described
Vocabulary Checklist (396 items)	Checklist organized into 19 semantic categories; response options are understands or understands and says	Words Understood; Words Produced
CDI: Words and Gestures (Part II—Actions and Gestures)		
First Communicative Gestures (12 items)	Checklist of intentional gestures	Early Gestures; Total Gestures
Games and Routines (6 items)	Checklist of games the child plays, such as pattycake or peekaboo	Early Gestures; Total Gestures
Actions with Objects (17 items)	Checklist of actions the child is able to perform, such as brushing teeth, combing hair, or eating with a spoon or fork	Later Gestures; Total Gestures
Pretending to Be a Parent (13 items)	Checklist of actions the child sometimes performs with stuffed animals or toys, such as putting it to bed or talking to it	Later Gestures; Total Gestures
Imitating Other Adult Actions (15 items)	Checklist of actions the child might try to imitate, such as cleaning with a broom, vacuuming, or washing dishes.	Later Gestures; Total Gestures
CDI: Words and Sentences (Part I--Words Children Use)		
Vocabulary Checklist (680 items)	Checklist of words the child can say, organized into 22 semantic categories	Words Produced
How Children Use Words (5 items)	Questions on the child's use of language for past, future, and absent objects and people	Not described
CDI: Words and Sentences (Part II--Sentences and Grammar)		
Word Endings/Part I (4 items)	Questions about the child's use of language to refer to past, future, and absent objects and people that differ from questions in the How Children Use Words subtest; for example, the subtest includes questions about how the child uses the possessive	Not described
Word Forms (25 items)	Checklist of irregular plural nouns and irregular past tense verbs	Word Forms
Word Endings/Part II (45 items)	Checklist of over-regularized nouns and verbs	Word Endings/Part II
Combining (1 item)	Question on whether the child can combine words into sentences	
Examples (1 item)	Request for parent to provide up to 3 of the longest sentences uttered by the child	Mean Length of the 3 Longest Sentences (M3L)
Complexity (37 items)	Parents select one from a pair of sentences contrasting in complexity to indicate how their child currently speaks	Complexity

Source: Fenson et al. User's Guide and Technical Manual, 2007.

not been collected. For the short forms, the correlations between the Infant (Level 1) short and long forms were 0.98 on Words Understood and 0.97 on Words Produced (Fenson et al. 2007). The overall correlations between the Toddler (Level 2) short forms (A and B) and the long form were each 0.99.

Other Languages: A Mexican Spanish version of the MacArthur-Bates CDI (the CDI: Words and Gestures and CDI: Words and Sentences forms) is available and called the Inventarios. The Spanish CDIs (and short forms) and the manual were published in 2003. The Inventarios were normed on more than 2,000 children. The CDI-III is currently not available in Spanish. The infant and toddler CDI forms have also been adapted in various languages such as Arabic, French, Finnish, Mandarin, Korean, and Malay, among others.

Uses of Information: The MacArthur-Bates CDI screens for delays in language development and to identify problematic skills. The developers also note that the CDI can help formulate intervention strategies and evaluate treatment outcomes.

Methods of Scoring: Scoring may be performed manually or electronically. The User's Guide and Technical Manual provide instructions for manual scoring. Scoring the inventories involves counting the number of marked items or affirmative responses by subtest. Thus, within an inventory, several subtests are combined to create a variety of composite scores (see Table 1). For the CDI: Words and Gestures inventory, the assessor determines five potential raw scores. In Part I's Vocabulary Checklist, for each of the 22 semantic subcategories, items marked "understands" yield the Words Understood score, and those marked "understands and says" yield the Words Produced score (each has a maximum score of 396). Items marked "yes" in Part II, First Communicative Gestures and Games and Routines, are summed to yield the Early Gestures score (maximum score of 18) while those marked "yes" in the other Part II subtests of Actions with Objects, Pretending to Be A Parent, and Imitating Other Adult Actions make up the Later Gestures score (maximum score of 45). The Early Gestures and Later Gestures scores are summed for a Total Gestures score.

For the CDI: Words and Sentences inventory, the items within each subtest are also summed to provide five potential raw scores (see Table 1). In Part I, the assessor calculates the Words Produced raw score by counting items marked as "says" in each of the 19 semantic subcategories of the Vocabulary checklist (maximum score of 680). Additional subtests such as Word Endings/Part I, Word Forms, and Word Endings/Part II are each individually scored by counting all items marked "sometimes" and "often" and computing total scores for each. The Examples subtest is scored by calculating the number of morphemes in each of the three example sentences and obtaining an M3L score (mean length of three longest sentences), instructions for which are provided in the User's Guide and Technical Manual. The Complexity subtest is scored by counting the number of items marked in the more complex of the two alternatives provided, yielding the Complexity score (maximum score of 37).

Use of tables in the User's Guide permits the conversion of raw scores into gender- and age-specific percentile rankings. An automated, free CDI scoring program is available at http://www.sci.sdsu.edu/cdi/. The program scores the English and Spanish forms as well as the short forms, calculates percentiles by using the raw scores, and generates child reports and parent letters.

The CDI-III is scored by computing raw total scores for each of its three subtests: (1) Vocabulary checklist (maximum score 100), (2) Sentences (maximum score 12), and (3) Using Language (maximum score 12). Assessors then convert the raw total scores into percentiles for comparison with the norming tables provided in the User's Guide and Technical Manual.

Interpretability: The User's Guide and Technical Manual provide instructions for interpreting the results. The normed percentile ranking allows the infant's or toddler's performance to be compared to that of other infants or toddlers. For the CDI: Words and Sentences inventory, the authors noted a ceiling effect for the Animal Sounds category within the Vocabulary Checklist but did not elaborate. In addition, the manual provides normed percentile information for 3-year-olds on the CDI-III. Even though the measure may be self-administered and scored by using the free downloadable automated program, the manual recommends that either a clinician or a researcher interpret the results.

Reliability:[2]
(1) Internal consistency reliability: Cronbach's alpha coefficients for scores from the CDI: Words and Gestures inventory for Words Produced and Words Understood were 0.95 and 0.96, respectively. For the Vocabulary Checklist, authors provided information only for semantic subcategories with coefficients below 0.70 (2 of 19): Words about Time (0.65 for both Words Produced and Words Understood) and Question Words (0.68 for Words Produced and 0.56 for Words Understood). The Total Gestures score had a reliability estimate of 0.88. In addition, scores for the Infant short form had a Cronbach's alpha of 0.97.

Cronbach's alpha coefficients for scores of the CDI: Words and Sentences inventory were 0.86 for the Words Produced scores and 0.95 for the Complexity scores (analyzed by using bound morphemes, functor words, and complex sentences). For the Vocabulary Checklist, authors provided information only for semantic subcategories with coefficients below 0.70 (2 of 22): Sound Effects and Animal Sounds (0.65) and Connecting Words (0.68). Scores for the Toddler short Forms A and B each demonstrated an internal consistency alpha of 0.99.
(2) Test-retest reliability: For CDI: Words and Gestures, on a sample of 137 children, the test-retest correlations were in the 0.80s for both Words Produced and Words Understood, with an average interval of 1.4 months between the first and second administrations. The authors do not provide age breakdowns but note that the correlation decreased to 0.61 for children assessed at 12 months. The Infant short form yielded test-retest reliability estimates of 0.88 for Words Understood and 0.90 for Words Produced, based on a two-week interval.

For CDI: Words and Sentences, on a sample of 216 children, the correlation was 0.95 for Words Produced, with all correlations above 0.90 across all ages and an average interval of 1.4 months between first and second administrations. The authors did not provide a breakdown by age. The test-retest reliability estimates were 0.74 and 0.93 for Words Understood for the Toddler short Forms A and B, respectively, with a two-week interval.

No test-retest reliability information was described for scores from the CDI-III.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: Not applicable.

Validity Evidence:
The authors drew items within each subtest from the developmental literature and used parent suggestions in response to earlier versions of the assessment. Major domains for infants include Words Understood, Words Produced, and Actions.

Construct/Concurrent validity: The authors correlated scores from the CDI with several language measures. For CDI: Words and Gestures, the authors correlated scores from Words Produced with scores from the Language Sample NDW (Number of Different Words), the Preschool Language Scale (PLS-Revised), the Peabody Picture Vocabulary Test–III (PPVT–III), and the Reynell Developmental Language Scales (RDLS) Expressive subtest, with correlations ranging from 0.52 to 0.82 between scores. For the CDI: Words and Sentences, the authors used the same measures as above as well as the Bayley Scales of Infant Development (Second Edition) Expressive Language subtest, the Expressive One-Word Picture Vocabulary Test, and Sequenced Inventory of Communication Development–Revised (SICD-R), with correlations ranging from 0.40 to 0.88 between scores. Lastly, for the CDI: Words and Gestures, the authors correlated scores from Words Understood to the Index of Productive Syntax, the PPVT–III, the RDLS Receptive, and the Language Sample NDW, with correlations ranging from 0.51 to 0.87.

Scores from the CDI-III correlated at 0.63 with the PLS-3 total score, 0.58 with the PLS-3 Auditory Comprehension Score, and 0.47 with the PLS-3 Expressive Communication Score for a sample of 19 children (36 and 37 months). The authors did not provide information on CDI-III subtests. Correlations for scores of the PPVT-R with the CDI-III Vocabulary Checklist, Sentences, and Using Language subtests were 0.50, 0.45, and 0.63, respectively, for a sample of 22 children (36 to 39 months). Two separate studies correlated scores from the CDI-III with scores from the McCarthy Scales, with correlations ranging from 0.44 to 0.62 in the first study of 85 32- to 40-month-olds and from 0.52 to 0.56 in the second study of 113 3-year-olds.

Predictive validity: Authors investigated predictive validity by correlating scores of the CDI forms with themselves, respectively, by scale, with a six-month interval between the first and second administrations.

The CDI: Words and Gestures six-month correlation between scores by subtest were 0.38 for vocabulary production, 0.44 for vocabulary comprehension, and 0.44 for total gestures, using a sample of 62 children (age 8 to 10 months at Time 1). A correlation of 0.69 was observed for Words Produced for a sample of 217 children age 10 to 16 months at Time 1 and age 16 to 25 months at Time 2.

Separate correlations computed to control for age showed a significant decrease in correlation at 12 months to 0.38, which, the authors noted, may be attributable to developmental transitions that occur at 12 to 13 months of age. The CDI: Words and Sentences correlated with itself six months later at 0.71 for Words Produced and 0.62 for Complexity scores based on a sample of 228 children (16 to 24 months at Time 1).

Bias Analysis: No information available.

Training Support: No information available.

Adaptations/Special Instructions for Individuals with Disabilities: The manual cautions against using the CDI with developmentally delayed children whose chronological age exceeds the inventory's upper limits.

Alternate Forms: None.

Previous Version: The original edition of the MacArthur Communicative Development Inventories was published in 1992 and consisted of two inventories: Words and Gestures and Words and Sentences. The updated edition of the MacArthur-Bates CDIs adds the CDI-III, an extension to capture information on children age 30 to 37 months. The norming data for the CDI: Words and Gestures were expanded to include 17- and 18-month-olds. Additional information on administration, interpretation, and scoring procedure options was added.

NCEE or REL Study Use:[3] Evaluating the Impact of the Program for Infant/Toddler Care

[1] This rating refers to the reliability for total test scores or scores commonly reported. Individual subtests encompassed some rating below the 0.70 level (see Reliability).

[2] Reliability and validity information was calculated by using specific scores highlighted in the User's Guide and Technical Manual. Not every score described in Table 1 was used in these calculations. The relevant sections of the profile list only those scores described by the authors for the corresponding calculations.

[3] See Table F.1 for web address.

References:

Fenson, Larry, Virginia A. Marchman, Donna J. Thal, Philip S. Dale, J.S. Reznick, and Elizabeth Bates. *MacArthur-Bates Communicative Development Inventories, User's Guide and Technical Manual, Second Edition*. Baltimore: Paul H. Brookes Publishing Co., 2007.

Fenson, Larry, Steve Pethick, Connie Renda, and Jeffrey L. Cox. "Short-Form Versions of the MacArthur Communicative Development Inventories." *Applied Psycholinguistics*, vol. 21, no. 1, 2000, pp. 95-115.

Kisker, Ellen E., Kimberly Boller, Charles Nagatoshi, Christine Sciarrino, Vinita Jethwani, Teresa Zavitsky, Melissa Ford, and John M. Love. "Resources for Measuring Services and Outcomes in Head Start Programs Serving Infants and Toddlers." Washington, DC: Mathematica Policy Research, 2003.

METACOGNITIVE AWARENESS OF READING STRATEGIES INVENTORY (MARSI), 2002

Authors: Kouider Mokhtari and Carla Reichard		**Type of Assessment:** Student self-report **Domain:** Reading (strategies)
Publisher: Unpublished; items and scoring rubric available in Mokhtari and Reichard (2002) and Mokhtari et al. (2008a)		**Grade/Age Range:** Grade 6 through college **Administration Interval:** As frequently as desired
Material, Training, and Scoring Costs: No costs noted		**Personnel and Training Requirements** **Credentials Required for Use:** No special qualifications required **Personnel for Administration:** Individual with basic clerical skills and some training **Training for Administration:** Self-training (< 1 hour)
Languages: English, Arabic, French, Spanish, and Hungarian		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (under $100) **Time to Administer:** 10 to 12 minutes **Ease of Administration and Scoring:** 2 (self-administered or administered and scored by someone with basic clerical skills) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The MARSI is a self-report tool that assesses students' awareness and use of reading strategies when reading academic materials. The inventory, designed for students in grade 6 through college, consists of 30 items across three subscales: Global Reading Strategies, Problem-Solving Strategies, and Support Reading Strategies. The average administration time is between 10 and 12 minutes. The Global Reading subscale (13 items) includes strategies such as previewing the material and predicting what the text will discuss. The Problem-Solving subscale (8 items) assesses students' use of strategies such as rereading or checking their understanding when reading difficult passages. The Support Reading subscale (9 items) involves students' use of strategies to facilitate reading such as taking notes or referring to a dictionary. Based on a five-point scale, the MARSI measures how frequently students use the 30 strategies: "I never or almost never do this;" "I do this only occasionally;" "I sometimes do this (about 50 percent of the time);" "I usually do this;" or "I always or almost always do this." An adapted version of the MARSI, the Survey of Reading Strategies, is available to assess the English reading strategies of English as a Second Language students.

Other Languages: The inventory has been translated into Arabic, French, Spanish, and Hungarian. These versions are available from the authors upon request (Mokhtari et al. 2008a).

Uses of Information: The MARSI assesses how frequently students use reading strategies to assist in comprehending academic materials. The authors state that students may use the results of the inventory to determine if they could use additional strategies to enhance their reading comprehension skills. Teachers may use the inventory to assess and monitor their students' reading strategies and comprehension processes and to guide and individualize instruction. One study used the MARSI to assess whether students use different strategies when reading academic material versus reading for pleasure (Mokhtari and Reichard 2008a). In addition, the MARSI may assess the effectiveness of interventions designed to increase students' awareness and use of reading strategies. The authors caution that the MARSI should be used only as a supplemental tool to assess students' reading comprehension, particularly given its self-report design. It should not replace existing reading assessments.

Methods of Scoring: Based on a five-point scale, the MARSI measures how frequently students use the 30 strategies. A total inventory score and separate scores for each of the three subscales may be computed. The total score is the sum of all the item responses. Students may score their own inventories by using the accompanying scoring rubric.

Interpretability: The authors have created three performance categories for the use of reading strategies. Students with a mean score of 3.5 or higher are categorized as high users, and students with a mean score of 2.4 or lower are categorized as low users. The categories are based on the performance of the 443 students who were administered the final version of the MARSI.

Reliability:
(1) Internal consistency reliability: Based on data from the final sample of 443 students in grades 6 through 12, Cronbach's alpha coefficients for scores ranged from 0.79 to 0.92 across subscales and from 0.86 to 0.93 across grade levels, with a reliability estimate of 0.89 for scores for the entire sample (Mokhtari and Reichard 2002).
(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: Not applicable.

Validity Evidence:
The authors based the items in the inventory on existing reading strategy assessments and a review of the reading research literature. An expert panel eliminated 40 redundant items from an initial pool of 100 potential items. The 60 remaining items were piloted with a sample of 825 students in grades 6 through 12 in 10 school districts across 5 Midwestern states. Fifty-two percent of the students were White, 19 percent were American Indian, 4 percent were Asian, 6 percent were Black, 7 percent were Hispanic, and 11 percent described themselves as Other. In addition to completing the inventory, the students provided feedback such as whether any of the items were unclear or confusing. Exploratory factor analysis identified the inventory's three subscales. The 60 items were examined relative to their discrimination power, redundancy, and factor loading, with 30 items retained in accordance with factor loadings of at least 0.30 for at least one factor. The final items were included in a factor if their factor loadings were at least 0.20 for a given factor. The expert panel approved the 30 items.

Construct/Concurrent validity: The 30-item inventory was administered to a sample of 443 students in grades 6 through 12 demographically similar to the initial pilot sample described above. A factor analysis was conducted, and the three previously identified subscales explained 29.7 percent of the total variance (Mokhtari and Reichard 2002). One study of 51 high school students looked at whether students' general report of reading strategies differed from the strategies that they reported actually using after reading a textbook chapter. The study found that students reported using fewer strategies after they were asked to reflect on what strategies they had actually used when reading an academic text versus the strategies they initially reported using when thinking more generally about their reading strategies (Mokhtari et al. 2008b).

The authors utilized a one-way analysis of variance (ANOVA) and post hoc Ryan-Einot-Gabriel-Welsh multiple range tests to compare students' self-reported reading ability with their use of reading strategies. The results indicated that students who rated their reading ability as excellent used significantly more global and problem-solving reading strategies than students who rated their reading ability as average or not so good (Mokhtari and Reichard 2002).

In a study involving 65 grade 11 students, a repeated measures ANOVA was conducted to determine if students' reported strategy use differed by gender or reading ability. The study found no significant differences, but females and less skilled readers reported using more reading strategies (Mokhtari and Reichard 2008a). Another study comparing the use of reading strategies between 10 high-achieving and underachieving gifted grade 8 students found no significant difference (Berkowitz and Cicchelli 2004).

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: No special training required.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: None.

Previous Version: None.

NCEE or REL Study Use:[1] Impact of the Thinking Reader Software Program on Grade 6 Reading Comprehension, Vocabulary, Strategies, and Motivation (REL-Northeast & Islands)

[1] See Table F.1 for web address.

References:

Berkowitz, Esther, and Terry Cicchelli. "Metacognitive Strategy Use in Reading of Gifted High Achieving and Gifted Underachieving Middle School Students in New York City." *Education & Urban Society*, vol. 37, no. 1, 2004, pp. 37-57.

Mokhtari, Kouider, and Carla A. Reichard. "Assessing Students' Metacognitive Awareness of Reading Strategies." *Journal of Educational Psychology*, vol. 94, no. 2, 2002, pp. 249.

Mokhtari, Kouider, and Carla A. Reichard. "The Impact of Reading Purpose on the Use of Reading Strategies." In *Reading Strategies of First- and Second-Language Learners: See How They Read*, edited by Kouider Mokhtari and Ravi Sheorey. Norwood, MA: Christopher-Gordon Publishers, Inc., 2008a.

Mokhtari, Kouider, and Carla A. Reichard. "Using Rasch Analysis to Calibrate Students' Metacognitive Awareness and Use of Reading Strategies." In *Reading Strategies of First- and Second-Language Learners: See How They Read*, edited by Kouider Mokhtari and Ravi Sheorey. Norwood, MA: Christopher-Gordon Publishers, Inc., 2008b.

Mokhtari, Kouider, Ravi Sheorey, and Carla A. Reichard. "Measuring the Reading Strategies of First- and Second-Language Readers." In *Reading Strategies of First- and Second-Language Learners: See How They Read*, edited by Kouider Mokhtari and Ravi Sheorey. Norwood, MA: Christopher-Gordon Publishers, Inc., 2008a.

Mokhtari, Kouider, Carla A. Reichard, and Ravi Sheorey. "Metacognitive Awareness and Use of Reading Strategies among Adolescent Readers." In *Reading Strategies of First- and Second-Language Learners: See How They Read*, edited by Kouider Mokhtari and Ravi Sheorey. Norwood, MA: Christopher-Gordon Publishers, Inc., 2008b.

MOTIVATION FOR READING QUESTIONNAIRE (MRQ), 1997

Authors: Allan Wigfield and John T. Guthrie		**Type of Assessment:** Student self-report **Domain:** Approaches toward learning/motivation (reading-specific)
Publisher: Unpublished; items and response format listed in Wigfield and Guthrie (1997) and Baker and Wigfield (1999)		**Grade Range:** Grades 3 through 6 **Administration Interval:** None described
Material, Training, and Scoring Costs: No costs noted		**Personnel and Training Requirements** **Credentials Required for Use:** Not specified **Personnel for Administration:** Individual with basic clerical skills and some training **Training for Administration:** Self-training (< 1 hour)
Languages: English		**Alternate Forms:** No
Representativeness of Norming Samples: No norming sample		**Summary** **Initial Material Cost:** 1 (< $100) **Time to Administer:** 15 to 20 minutes **Ease of Administration and Scoring:** 1 (not described) **Reliability:** 2 (all or mostly under 0.70) **Predictive Validity:** Available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The MRQ is a 54-item student self-report questionnaire designed to assess dimensions of reading motivation. It has been used with samples of students in grades 3 through 6. It consists of 11 Reading Motivation Scales designed to assess perceptions of self-efficacy, intrinsic and extrinsic motivation, and social motivation for reading: Reading Efficacy (three items), Challenge (five items), Curiosity (six items), Involvement (six items), Importance (two items), Recognition (five items), Grades (four items), Social (seven items), Competition (six items), Compliance (five items), and Reading Work Avoidance (four items). The measure does not yield an overall summary score. The response format is a 4-point scale (1 = very different from me, 2 = a little different from me, 3 = a little like me, 4 = a lot like me). The MRQ takes 15 to 20 minutes to complete and may be group-administered.

Other Languages: None.

Uses of Information: The MRQ is designed for research and evaluation use. It has been used in studies investigating associations between students' reading motivation and demographic characteristics, reading outcomes, and the effectiveness of reading curricula and incentive programs designed to increase the amount of time students spend reading.

Methods of Scoring: Published descriptions of the MRQ do not include specific scoring information (Baker and Wigfield 1999; Wigfield and Guthrie 1997). In these studies, the authors computed scale scores for the 11 scales by summing scores across scale items and computing means and standard deviations. Wigfield and Guthrie (1997) also combined scores on the Efficacy, Curiosity, and Involvement scales to form an Intrinsic composite score and on the Recognition, Grades, and Competition scales to form an Extrinsic composite score. The authors noted the composites were based on factor analyses and "theoretical distinctions in the motivation literature" (p. 425).

Interpretability: Guidelines for MRQ score interpretation are not readily available. In general, the interpretability of data derived from the MRQ is unclear given that independent investigations of its construct validity yielded alternative factor structures (Watkins and Coffey 2004; see construct/concurrent validity below).

Reliability:
(1) Internal consistency reliability: Wigfield and Guthrie (1995) reported internal consistency reliabilities of 0.70 or greater for 5 of 11 "factor-based" scales (ranging from 0.43 to 0.81). Baker and Wigfield (1999) also conducted a factor analysis of 11 theory-derived subgroups of items and reported alpha coefficients of 0.70 or greater for 5 of 11 scales (ranging from 0.55 to 0.76). In exploratory factor analyses conducted with two samples, Watkins and Coffey (2004) found different 8-factor solutions for each sample. In the first analysis, alpha coefficients for the 8 factors ranged from 0.60 to 0.75 (with 4 of 8 factors $\geq$ 0.70); in the second analysis, alpha coefficients ranged from 0.54 to 0.80 (with 5 of 8 factors $\geq$ 0.70).
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: Not applicable.

Validity Evidence:
In developing the MRQ, Wigfield and Guthrie (1995) reviewed motivation theory and research and identified constructs hypothesized to relate to reading activity. They also reviewed student interviews and observations of classroom reading instruction (Guthrie et al. 1996). The MRQ focuses on three sets of motivation-related constructs: (1) beliefs about efficacy; (2) individuals' reasons for performing tasks (including intrinsic and extrinsic motivation, achievement goals, and value attached to achievement); and (3) social aspects of motivation.

Construct/Concurrent validity: Wigfield and Guthrie (1997) and Baker and Wigfield (1999) conducted factor analyses of selected groups of MRQ items with two convenience samples (N = 105 and N = 650) of economically and ethnically diverse students in grades 4 through 6 in predominantly urban, mid-Atlantic elementary schools participating in reading incentive programs. In both studies, sample size limitations precluded factor analysis of the entire set of MRQ items; the authors instead performed separate factor analyses on subgroups of items theorized to represent the 11 reading motivation dimensions. In both investigations, the authors concluded that the MRQ measures 11 dimensions of reading motivation.

In a study investigating the structural validity of the MRQ with two samples (N = 328 and N = 735) of suburban, socioeconomically diverse, mostly White students in grades 3 through 5 in mid-Atlantic and southwestern states, Watkins and Coffey (2004) found factor structures that did not replicate the factor structure identified in earlier studies (Wigfield and Guthrie 1997; Baker and Wigfield 1999). They contended that the earlier investigations were marked by sample and methodological issues, including small nonrepresentative samples, flaws in factor analysis approaches, and time limits imposed on respondents that may have biased results by selecting for fluent readers. Watkins and Coffey (2004) argue that ". . . neither the MRQ nor its scales should be used as dependent variables in reading motivation research . . . or as measures of affective change in high-stakes educational evaluations" (p. 117).

Correlations between scores on some MRQ scales and student self-reports of amount and breadth of reading ranged from 0.21 to 0.51 (Wigfield and Guthrie 1997). Baker and Wigfield (1999) reported similar correlations between reading motivation and reported reading activity, ranging from 0.14 to 0.51. Only 1 of the 11 motivation scales correlated with reading achievement—scores on the Reading Work Avoidance scale correlated negatively with scores on two standardized tests of reading achievement (*r*s = -0.26 and -0.24). The Reading Work Avoidance, Compliance, Grades, and Recognition scales demonstrated correlations ranging from -0.13 to 0.21 with ratings of students' responses to questions about short stories on a curriculum-based reading performance assessment. Wigfield and Guthrie (1997) also reported that students with higher intrinsic motivation scores read nearly three times as many minutes per day as students with lower intrinsic motivation scores.

Some MRQ scale scores differ by gender and grade level. Wigfield and Guthrie (1997) reported that girls scored higher than boys on the Efficacy, Importance, and Social scales, whereas boys outscored girls on the Competition scale. Baker and Wigfield (1999) found higher scores for girls on all of the scales except Competition and Reading Work Avoidance, which did not vary by gender. With respect to grade level, fourth graders outscored fifth graders on the Efficacy, Recognition, and Social scales in fall of the school year, but not in spring (Wigfield and Guthrie

1997), and fifth graders outscored sixth graders on the Social and Recognition scales (Baker and Wigfield 1999).

Predictive validity: Wigfield and Guthrie (1997) reported correlations ranging from 0.21 to 0.36 between scores on some MRQ scales assessed in fall of the school year and the amount of outside-of-school reading students logged across the school year.

Bias Analysis: No information available.

Training Support: No information available.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: None.

Previous Version: None.

NCEE or REL Study Use:[1] Impact of the Thinking Reader Software Program on Grade 6 Reading Comprehension, Vocabulary, Strategies, and Motivation

[1] See Table F.1 for web address.

References:

Baker, Linda, and Allan Wigfield. "Dimensions of Children's Motivation for Reading and Their Relations to Reading Activity and Reading Achievement." *Reading Research Quarterly*, vol. 34, 1999, pp. 452-457.

Guthrie, John T., Karen McGough, and Allan Wigfield. "Measuring Reading Activity: An Inventory." (Instructional Resource No. 4). Athens, GA: National Reading Research Center, 1994.

Guthrie, John T., Peggy Van Meter, Ann Dacey McCann, Allan Wigfield, Lois Bennett, Carol C. Poundstone, Mary Ellen Rice, Frances M. Faibisch, Brian Hunt, and Ann M. Mitchell. "Growth in Literacy Engagement: Changes in Motivations and Strategies during Concept-Oriented Reading Instruction." *Reading Research Quarterly*, vol. 31, no. 3, 1996, pp. 306-332.

Watkins, Marley W., and Debra Y. Coffey. "Reading Motivation: Multidimensional and Indeterminate." *Journal of Educational Psychology*, vol. 96, no. 1, 2004, pp. 110-118.

Wigfield, Allan, and John T. Guthrie. "Dimensions of Children's Motivations for Reading: An Initial Study." (Research Report No. 34). Athens, GA: National Reading Research Center, 1995.

Wigfield, Allan, and John T. Guthrie. "Relations of Children's Motivation for Reading to the Amount and Breadth of their Reading." *Journal of Educational Psychology*, vol. 89, no. 3, 1997, pp. 420-432.

Wigfield, Allan, John T. Guthrie, and Karen McGough. "A Questionnaire Measure of Children's Motivations for Reading." (Instructional Resource No. 22). Athens, GA: National Reading Research Center, 1996.

Wigfield, A., Kathleen Wilde, Linda Baker, Sylvia Fernandez-Fein, and Deborah Scher. "The Nature of Children's Motivations for Reading and Their Relations to Reading Frequency and Reading Performance." (Reading Research Report No. 63). Athens, GA: National Reading Research Center, 1996.

NORTHWEST EVALUATION ASSOCIATION (NWEA) MEASURES OF ACADEMIC PROGRESS (MAP) AND ACHIEVEMENT LEVEL TESTS (ALT), 2003

Authors: Northwest Evaluation Association		**Type of Assessment:** Individually administered adaptive assessment (MAP) and group-administered assessment (ALT) **Domain:** Reading , language arts/language proficiency, mathematics, and science
Publisher: Northwest Evaluation Association 503-624-1951 http://www.nwea.org/		**Grade/Age Range:** Grades 2 through 11 **Administration Interval:** Up to four times in an academic year (MAP)
Material, Training, and Scoring Costs: Not available		**Personnel and Training Requirements** **Credentials Required for Use:** Not specified **Personnel for Administration:** Self- or computer-administered, computer-scored (MAP); individual with basic clerical skills with some training (ALT); **Training for Administration:** Minimal (1 to 2 hours)
Languages: English; Spanish audio for mathematics		**Alternate Forms:** Yes with adaptive selection of items each administration; MAP may be administered up to four times a year; ALT interval not described
Representativeness of Norming Sample: None described; norms are available from 2005 and 2008; however, publications describing the norming samples are not publicly available.		**Summary** **Initial Material Cost:** To be determined upon negotiation with the publisher **Time to Administer:** About 50 to 75 minutes **Ease of Administration and Scoring:** 2 (self-administered or administered and scored by someone with basic clerical skills) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Northwest Evaluation Association (NWEA) offers two assessments for students in grades 2 through 11: the Measures of Academic Progress (MAP), which is an individually administered, computerized, adaptive assessment; and the Achievement Level Tests (ALT), which is a group-administered pencil-and-paper assessment. (A separate adaptive assessment is also available for students in kindergarten through grade 2.) Both ALT and MAP assessments can assess students in one or more of four domains: Mathematics, Reading, Science, and Language Usage. Both assessments are untimed but typically take between 50 and 75 minutes for the average student and feature between 40 and 50 items. Both assessments draw items from a central item bank developed by NWEA. The assessments are flexible and may be tailored to the specific needs of individual organizations. Each domain includes goal areas which consist of at least seven items, and each test typically involves four to eight goals. In turn, each goal embodies five or six subgoals or objectives. Details of the coverage of the domains are not provided, but one example illustrates that reading contains goals of word meaning and comprehension and that use of context clues or use of synonyms, antonyms, and homonyms are subgoals under the word meaning goal. The MAP assessment draws items from a pool of 1,200 to 2,400 pre-calibrated items (depending on domain) during the assessment and adaptively administers subsequent items based on a student's performance on previously presented items. The level of difficulty for the initial item is determined by a student's previous results or grade level. The ALT assessments are a series of paper-and-pencil tests designed for students at different achievement levels within the domain being tested, with items drawn from the NWEA item banks based on an organization's predetermined goals or needs. Each test in the ALT series partially overlaps with the previous and subsequent tests in terms of difficulty while a previous level includes items that are easier on average than the items of the subsequent level. Determination of the appropriate ALT test level is based, if possible, on a student's previous three years of ALT results; otherwise, a short locator test may be given to determine the student's level of performance on the scale before selecting the most appropriate test level.

Other Languages: A Spanish audio version of the MAP mathematics test is available for purchase. It presents questions to students in spoken, formal Castilian Spanish. No information is provided as to how the translation was created or how comparable it is to the English wording. No information is provided on whether the audio version was normed with a sample of Spanish listeners.

Uses of Information: The MAP and ALT assessments may be used to measure the achievement level and growth of students in any of four domains: Mathematics, Reading, Science, and Language Usage. The developer also notes that scores may be used for course placement, parent conferences, and district-wide testing, for identifying a student's appropriate instructional level, and for screening students for placement in special programs.

Methods of Scoring: The proctor hand-scores locator tests for the ALT by using the scoring key provided with the tests. Either the assessor or NWEA scores the ALT by using Scoring and Reporting Software. The software also invalidates a student's score and recommends retesting if the percentage of items answered correctly was equal to or less than the percentage correct obtained by guessing plus 5 percent; if the percentage correct is greater than or equal to 95 percent; if the student answered less than half the items; or if the standard error of measurement

(SEM) is greater than 5.3 Rasch unit (RIT) points.[1] A computer scores the MAP assessment iteratively during administration, with the score available immediately after completion of the assessment. A student's score is invalidated and the student retested if he or she took less than six minutes to complete the assessment or if the standard error of measurement is greater than 5.5 RIT points (unless the score is greater than 240 RIT) or less than 1.0 RIT point. The scoring software provides the following score information: RIT score, SEM, RIT range, performance on each of the assessment's goal areas, percentile rank, percentile range, and a Lexile score (for reading assessments only). The scoring software can produce reports at the student, class, grade, school, or district level.

Interpretability: NWEA provides online and paper resources, such as annotated sample reports, to help interpret results and assigns a contact person trained in interpretation to each test site. NWEA also provides interpretation of normal performance for percentile scores and organizes workshops to train educational agencies in interpreting and using the scores.

Reliability:
(1) Internal consistency reliability: NWEA calculated a marginal reliability coefficient, described as combined single index of measurement error estimated at different points on the achievement scale. The samples, taken from the NWEA norming studies from 1996 and 1999, generally exceeded 10,000 students in each grade level (except for grade 2, which typically had approximately 4,000 students). Reliability coefficients for scores of the ALT assessments ranged from 0.90 to 0.94 for Reading, from 0.93 to 0.95 for Mathematics, and from 0.89 to 0.93 for Language Usage. Reliability coefficients for scores of the MAP ranged from 0.93 to 0.95 for Reading, from 0.92 to 0.96 for Mathematics, and from 0.92 to 0.94 for Language Usage.
(2) Test-retest reliability: NWEA estimated test-retest reliability by administering alternate versions of the ALT or MAP assessments to the same students with an interval of 7 to 12 months, using norming samples from 1999 and 2002. Comparisons covered fall to spring, spring to fall, and spring to spring, and sample sizes for each grade generally exceeded 10,000 students (except for grade 2, which typically included between 4,000 and 6,000 students). Reliability coefficients for scores based on test-retest of the ALT ranged from 0.76 to 0.89 for Reading, from 0.70 to 0.93 for Mathematics, and from 0.77 to 0.90 for Language Usage. Reliability coefficients for alternate versions of the ALT and MAP across a similar time period ranged from 0.80 to 0.92 for Reading, from 0.77 to 0.94 for Mathematics, and from 0.88 to 0.92 for Language Usage. In addition, in another study, NWEA correlated the results of 4,883 grade 4 and 5 students who had taken the ALT assessments in spring and then the following fall; coefficients were 0.90 for Language Usage, 0.88 for Reading, and 0.89 for Mathematics. (Some researchers would interpret what the authors have described as test-retest reliability as evidence of predictive validity because of the longer testing interval.)
(3) Alternate form reliability: The MAP is an adaptive assessment that administers different items to students each time; thus, each administration is an alternate form of the previous administration such that test-retest reliability information provides evidence of reliability of the forms. No information is available for alternate form reliability of the ALT.
(4) Inter-rater reliability: No information available.

Validity Evidence:
During NWEA-conducted workshops, classroom teachers develop items drawn from various educational institutions. The teachers also suggest appropriate grade ranges for the items and

content categories for NWEA review. Items are field tested in one of three ways: a minitest of new items presented to students after the actual assessment; inclusion of the new items in the assessment; and administration of a special test of mostly new items to students not otherwise taking an assessment. Both adjusted point-biserial correlations and adjusted root mean square fit indices for the items undergo review to determine how well they perform for their respective scales. If an item performs well, it is added to the item bank. If an item performs poorly, it is revised and field tested again. If it still performs poorly, it is either excluded from the item bank or retested in a different grade. In addition, NWEA periodically reviews the item banks to determine if the scale has fluctuated or drifted over time by re-calibrating items several years after initial calibration.

Construct/Concurrent validity: Mathematics, Reading, and Language Usage ALT assessment scores correlated with the Stanford Achievement Test, Ninth Edition and ranged from 0.78 to 0.88. Mathematics, Reading, and Language Usage MAP scores correlated with the Iowa Test of Basic Skills and ranged from 0.74 to 0.84.

NWEA also correlated its assessment scores with scores from individual state assessments. Correlations ranged from 0.66 to 0.87 for Reading, from 0.72 to 0.90 for Mathematics, and from 0.60 to 0.85 for Language Usage. Mathematics and Reading scores correlated with the Arizona Instrument to Measure Standards and ranged from 0.69 to 0.80 for both the ALT and MAP. Mathematics, Reading, and Language Usage ALT assessment scores correlated with scores from the Colorado Student Assessment Program and ranged from 0.79 to 0.90 for the ALT assessment in 2002 and 0.84 to 0.92 for the ALT assessment in 2000. Mathematics and Reading MAP scores correlated with scores from the Illinois Standards Achievement Tests and ranged from 0.79 to 0.87. Mathematics, Reading, and Language Usage assessment scores correlated with scores from the assessment Indiana Statewide Testing for Educational Progress-Plus, ranging from 0.72 to 0.88 for the ALT and MAP in 2003 and from 0.74 to 0.90 for the ALT in 2000. Mathematics and Reading assessment scores correlated with scores from the Minnesota Comprehensive Assessment and Basic Skills Test and ranged from 0.77 to 0.85 for both the ALT and MAP. Mathematics and Reading MAP assessment scores correlated with scores from the Nevada Criterion Referenced Assessment, ranging from 0.76 to 0.86. Mathematics and Reading ALT assessment scores correlated with scores from the Palmetto Achievement Challenge Tests, ranging from 0.70 to 0.87. Mathematics, Reading, and Language Usage MAP assessment scores correlated with scores from the Texas Assessment of Knowledge and Skills and ranged from 0.66 to 0.82. Mathematics and Reading ALT assessment scores correlated with scores from the Washington Assessment of Student Learning and ranged from 0.80 to 0.85. Mathematics, Reading, and Language Usage ALT assessment scores correlated with scores from the Wyoming Comprehensive Assessment System, ranging from 0.60 to 0.81. Sample sizes all exceeded 1,000 students for each grade and were as high as nearly 8,000 students. Students were generally in grade 3 through 10, except for analyses with scores from the Stanford Achievement Test, which included students in grade 2.

Predictive validity: A group of students' grade 9 ALT scores correlated with their grade 10 Washington Assessment of Student Learning scores. The correlation for Mathematics (N = 849) was 0.81, and the correlation for Reading (N = 1,003) was 0.75. In addition, NWEA correlated the scores for 3,677 grade 4 and 5 students who had taken the ALT in the spring and the MAP in the fall; coefficients were 0.83 for Language Usage and Reading and 0.85 for Mathematics. (As

noted, test-retest reliability was obtained over a 7- to 12-month testing interval; some researchers would consider the result as evidence of predictive validity. See test-retest reliability above for more information.)

Bias Analysis: As items undergo development, peer editors assess them for bias. NWEA also conducts bias review panels, in which a panel of stakeholders from a variety of racial and ethnic backgrounds reviews items. The panels either send back questionable items to the original author for revision or reject them for inclusion in the item bank.

Training Support: No information available.

Adaptations/Special Instructions for Individuals with Disabilities: Students with Individualized Education Plans (IEP) may be granted six types of accommodations: changes in timing or scheduling of the assessment; changes in how directions are presented; changes in how questions are presented; changes in how the student responds to questions; changes in the test setting; and changes in the references and tools provided during the assessment.

Alternate Forms: The MAP assessment is completely adaptive, such that different students receive different samples of items, while ensuring the comparability of assessment results across students. The MAP and the ALT draw on the same item bank. The MAP may be administered up to four times a year; no guidance is given on how frequently the ALT may be administered.

Previous Version: The item bank is continually updated.

NCEE or REL Study Use:[2] The Effects of Success in Sight as a School Improvement Intervention; Assessing the Impact of Collaborative Strategic Reading (CSR) on Reading Comprehension; The Impact of Professional Development Strategies on Teacher Practice and Student Achievement in Math

[1] Y RITs = (X logits * 10) + 200

[2] See Table F.1 for web address.

References:

Kingsbury, G. Gage. "An Empirical Comparison of Achievement Level Estimates from Adaptive Tests and Paper-and-Pencil Tests." Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April 2002.

Northwest Evaluation Association. "2008 Normative Data." Available at [http://www.nwea.org/assets/downloads/980/Normative%20Data%20Sheet_v2.pdf]. 2008.

Northwest Evaluation Association. "2005 Normative Data: Monitoring Growth in Student Achievement." Available at [http://docs.abileneschools.org/~deniseguy/Assessment/Site/Assessment_files/NormativeDataSheet.pdf]. 2005.

Northwest Evaluation Association. *Northwest Evaluation Achievement Level Tests (NWEA ALT) Mathematics, Reading, and Language Usage: Administration Guide for Paper-Pencil Tests*. Lake Oswego, OR: Northwest Evaluation Association, n.d.

Northwest Evaluation Association. "Reliability and Validity Estimates, NWEA Achievement Level Tests and Measures of Academic Progress." Lake Oswego, OR: Northwest Evaluation Association, 2004.

Northwest Evaluation Association. *Technical Manual for the NWEA Measures of Academic Progress and Achievement Level Tests*. Portland, OR: Northwest Evaluation Association, 2003.

PATTERNS OF ADAPTIVE LEARNING SCALES (PALS), 2000

Authors: Carol Midgley, Martin L. Maehr, Ludmila Z. Hurda, Eric Anderman, Lynley Anderman, Kimberley E. Freeman, Margaret Gheen, Avi Kaplan, Revathy Kumar, Michael J. Middleton, Jeanne Nelson, Robert Roeser, and Timothy Urdan		**Type of Assessment:** Student self-report (group-administered)[1] **Domain:** Approaches to learning/motivation
Publisher: The University of Michigan Michael Middleton 603- 862-7054 http://www.umich.edu/~pals/index.html		**Grade/Age Range:** Grades 4 through 9 **Administration Interval:** None described
Material, Training, and Scoring Costs: The manual, which includes student and teacher questionnaires, may be downloaded from the web site for free.		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2 (bachelor's degree with coursework in measurement and domain) **Personnel for Administration:** Individual with basic clerical skills with some training **Training for Administration:** Minimal (1 to 2 hours) Developers recommend administration of the student questionnaire by someone with knowledge of the confidentiality guidelines and administration practices described in the manual.[2] Interpreters of results should have completed coursework in survey design and quantitative research methods.
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (under $100) **Time to Administer:** 40 minutes **Ease of Administration and Scoring:** 3 (administered and scored by highly trained individual) **Reliability:** 3 (all at or above 0.70)[3] **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Patterns of Adaptive Learning Scales (PALS) assess perceptions of why individuals try to achieve. The student questionnaire is a group-administered self-report for students in grades 4 through 9. It includes five scales that may be used together or individually. The full questionnaire includes 106 items; students circle responses on paper forms. The five scales are (1) Personal Achievement Goal Orientations (14 items); (2) Perception of Teacher's Goals (12 items); (3) Perception of Classroom Goal Structures (14 items); (4) Academic-Related Perceptions, Beliefs, and Strategies (44 items); and (5) Perceptions of Parents, Home Life, and Neighborhood (22 items). The 5-point Likert-type scale responses are anchored at 1 (not at all true), 3 (somewhat true), and 5 (very true). Developers recommend that administration sessions last no longer than 40 minutes and note that they have administered the questionnaire over a two-day period.

The first three scales each have three subscales that measure goal perceptions along three lines: (1) Personal Mastery (i.e., concern with gaining competence or improving), (2) Performance-Approach (i.e., demonstrating competence compared to others), and (3) Performance-Avoidance (i.e., avoiding the appearance of incompetence compared to others). The Academic-Related Perceptions, Beliefs, and Strategies scale has eight subscales: (1) Academic Efficacy, (2) Academic Press (i.e., perception that teachers press students for understanding), (3) Academic Self-Handicapping, (4) Avoiding Novelty (i.e., avoiding new or unfamiliar work), (5) Cheating Behavior, (6) Disruptive Behavior, (7) Self-Presentation of Low Achievement (i.e., keeping peers from knowing how well students achieve in school), and (8) Skepticism about the Relevance of School for Future Success. The Perceptions of Parents, Home Life, and Neighborhood scale has four subscales: (1) Parent Mastery Goal (i.e., parents want students to develop competence), (2) Parent Performance Goal (i.e., parents want students to demonstrate competence), (3) Dissonance between Home and School (i.e., concern that home and school life differ), and (4) Neighborhood Space (i.e., able to find safe and enjoyable places in the neighborhood).

Other Languages: None.

Uses of Information: The PALS contributes to the body of research on achievement goal theory in that it clearly distinguishes among three types of goal perceptions—Personal Mastery, Performance-Approach, and Performance-Avoidance—when assessing individual perceptions about the purposes of achievement. PALS developers use this goal structure to examine the relationship between the learning environment and student motivation, affect, and behavior. More broadly, PALS research can yield recommendations for changes to the learning environment and guide school reform.

Methods of Scoring: Published descriptions of the PALS do not include specific scoring information (Midgley et al. 2000; Anderman and Midgley 2002), although the manual presents subscale means in its summary of descriptive statistics. The manual indicates three items on the Neighborhood Space subscale that require reverse coding of Likert-type responses. Information is not available on who conducts scoring.

Interpretability: For the Personal Achievement Goal Orientations, Perception of Teacher's Goals, and Perception of Classroom Goal Structures scales, high scores on each of the three respective subscales (Personal Mastery, Performance-Approach, and Performance-Avoidance) indicate high value for engaging in academic endeavors. Similarly, for Academic-Related Perceptions, Beliefs, and Strategies, high scores on two subscales (Academic Efficacy and Academic Press) indicate stronger perceptions toward engaging in academic endeavors. In contrast, high scores for the other six subscales (Academic Self-Handicapping, Avoiding Novelty, Cheating Behavior, Disruptive Behavior, Self-Presentation of Low Achievement, and Skepticism about the Relevance of School for Future Success) indicate stronger perceptions toward avoidance of engaging in academic endeavors. Finally, for the Perceptions of Parents, Home Life, and Neighborhood scale, high scores on three of the subscales (Parent Mastery Goal, Parent Performance Goal, and Neighborhood Space) indicate positive perceptions about competence and the neighborhood, whereas a high score on the Dissonance between Home and School subscale indicates stronger perceptions of concern about differences between home and school life. Guidelines for total PALS score interpretation are not readily available.

Reliability: Most reliability-related information pertains to earlier versions of the PALS student scales (see Previous Version). The manual does not indicate which reliability statistics were based on the 2000 PALS.
(1) Internal consistency reliability: Anderson and Midgley (2002) calculated Cronbach's alpha coefficients for scores from each subscale, with each subscale coefficient representing students from one grade between grades 5 and 9. Subscale coefficients ranged from 0.74 to 0.89 for the Personal Achievement Goal Orientations (grade 6), from 0.71 to 0.83 for the Perception of Teacher's Goals (grade 9), from 0.70 to 0.83 for the Perception of Classroom Goal Structures (grade 7), from 0.78 to 0.89 for Academic-Related Perceptions, Beliefs, and Strategies (grades 5, 6, and 7), and from 0.71 to 0.76 for the Perceptions of Parents, Home Life, and Neighborhood (grades 5 and 7). Ross et al. (2002) reported alpha coefficients for scores from the 1997 PALS Personal Achievement Goal Orientations subscales for two additional age groups, with coefficients from 0.79 to 0.82 for grade 4 students and from 0.70 to 0.85 for college students.

Ross et al. (2005) reported Cronbach's alpha coefficients to assess the variability across 103 coefficients from 30 research studies (mainly focused on students in grades 5 through 8) that used the PALS Personal Achievement Goal Orientations subscales before 1999. Average coefficients were 0.79, 0.79, and 0.81, respectively, for the Personal Mastery, Performance-Approach, and Performance-Avoidance subscales.
(2) Test-retest reliability: Developers conducted within- and across-grade correlations for the PALS Personal Achievement Goal Orientation subscales as students transitioned to the next school semester or year, respectively. Within-grade coefficients for grade 8 students (fall and spring) were 0.60, 0.61, and 0.54 for Personal Mastery, Performance-Approach, and Performance-Avoidance subscales, respectively. Across-grade coefficients for the same subscales for grade 8 and 9 students were 0.55, 0.58, and 0.39, respectively (Anderman and Midgley 2002). Within-grade coefficients for grade 5 students (fall and spring) were 0.63 for Personal Mastery and 0.61 for Performance-Approach. Across-grade coefficients for the same subscales for grade 5 and 6 students transitioning from elementary to middle school were 0.41 and 0.34, respectively.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: Not applicable.

Validity Evidence:
University of Michigan researchers developed the PALS between 1990 and 2000 to assess constructs associated with achievement goals. Developers refined the PALS for use in the Patterns of Adaptive Learning Study (Anderman and Midgley 2002), a large-scale longitudinal study in which questionnaires were administered to over 800 students in grades 5 through 9 (in 36 elementary, middle, and high schools) in four Michigan school districts between 1994 and 1999. Most validity-related information pertains to the 1997 version of the PALS student scales (see Previous Version for description). Information pertaining to the 2000 PALS is indicated when available.

Construct/Concurrent validity: Developers reported that they conducted exploratory factor analysis on PALS scales (data not available) (Anderman and Midgley 2002; Midgley et al. 1998). They removed three items (not specified) from the Personal Achievement Goal Orientations student scale that assessed intrinsic value or referenced specific behaviors. They conducted confirmatory factor analysis by examining the three goal perception factors (i.e., Personal Mastery, Performance-Approach, and Performance-Avoidance) and reported a goodness-of-fit Index (GFI) and adjusted goodness-of-fit index (AGFI) of 0.97 and 0.95, respectively. Developers also conducted confirmatory factor analysis on the same three goal perception factors for a new scale, the Perception of Classroom Goal Structure, which had a GFI of 0.96 and an AGFI of 0.94.

Several studies have assessed the relationship between the 1997 PALS Personal Achievement Goal Orientation subscales (i.e., Personal Mastery, Performance-Approach, and Performance-Avoidance) and other constructs and indicators. Midgley et al. (1998) reported Cronbach's alpha coefficients of 0.67 between Personal Mastery and the Task scale developed by Nicholls and colleagues (1989) and a coefficient of 0.63 between Performance-Approach and Nicholls and colleagues' Ego scale. Midgley et al. (1998) cited previous studies demonstrating that Personal Mastery goals were positively related to academic efficacy (cognition), adaptive learning strategies, and indices of affect. Studies showed mixed findings for the relationship between Performance-Approach goals and academic efficacy and learning strategies and no relation to affect. Performance-Avoidance goals were negatively related to academic efficacy, positively related to maladaptive learning strategies, and unrelated to affect.

Middleton and Midgley (1997) correlated the three PALS Personal Achievement Goal Orientation subscale scores and reported that Performance Mastery was not correlated with Performance-Approach ($r = 0.04$) or Performance-Avoidance ($r = 0.01$) but that Performance-Approach and Performance-Avoidance subscales were positively correlated ($r = 0.56$).

Predictive validity: No information available.

Bias Analysis: Midgley et al (1998) conducted analyses on three subscales within the 1997 PALS Personal Achievement Goal Orientation scale to detect differences by race and gender among elementary and middle school students. Using confirmatory factor analysis, the authors tested whether model parameters varied statistically across subgroups for the purpose of identifying a different fit of the three-factor (i.e., three subscales) model for White and Black students and for girls and boys. The authors demonstrated no model differences when comparing

White and Black students; however, the first three items of the Personal Mastery subscale had different error variances for girls and boys.

Training Support: The manual provides assessors with information on what to say to students who will complete the scale, such as explaining the confidentiality of responses and why some questions may sound similar to students.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: None.

Previous Version: Developers modified the 1997 PALS in 2000 (1) to decrease the number of items from 17 to 14 on the Personal Achievement Goal Orientations scale and change the wording of most items to reduce the focus on specific behaviors or interests that students exhibit or teachers encourage while learning and (2) to add three scales: the Perception of Classroom Goal Structure; Academic-Related Perceptions, Beliefs, and Strategies; and Perceptions of Parents, Home Life, and Neighborhood.

NCEE or REL Study Use:[4] The Effects of Classroom Assessment for Student Learning (CASL) on Student Achievement

[1] The Patterns of Adaptive Learning Scales also include a self-administered teacher questionnaire that captures level of agreement on three scales: Perceptions of the School Goal Structure for Students, Approaches to Instruction, and Personal Teaching Efficacy. For more information, see *Manual for the Patterns of Adaptive Learning Scales* (Midgley et al. 2000).

[2] The manual indicates that trained research assistants should administer the student questionnaire. M. Middleton reported, however, that training sessions are no longer offered (personal communication, January 12, 2009).

[3] Most reliability-related information pertains to earlier versions of the PALS student scales.

[4] See Table F.1 for web address.

References:

Anderman, Eric M., and Carol Midgley. "Methods for Studying Goals, Goal Structures, and Patterns of Adaptive Learning." In *Goals, Goal Structures, and Patterns of Adaptive Learning*, edited by Carol Midgley. Mahwah, NJ: Lawrence Erlbaum Associates, 2002.

Anderman, Eric M., Tim Urdan, and Robert Roeser. "The Patterns of Adaptive Learning Survey: History, Development, and Psychometric Properties." Paper presented at the Indicators of Positive Development Conference. Washington, DC: Child Trends, 2003.

Middleton, Michael, and Carol Midgley. "Avoiding the Demonstration of Lack of Ability: An Under-Explored Aspect of Goal Theory." *Journal of Educational Psychology*, vol. 89, no. 4, 1997, pp. 710-718.

Midgley, Carol, Avi Kaplan, Michael Middleton, Martin L. Maehr, Timothy Urdan, Lynley H. Anderman, Eric Anderman, and Robert Roeser. "The Development and Validation of Scales Assessing Students' Achievement Goal Orientation." *Contemporary Educational Psychology*, vol. 23, no. 2, 1998, pp. 113-131.

Midgley, Carol, Martin L. Maehr, Ludmila Z. Hurda, Eric Anderman, Lynley Anderman, Kimberley E. Freeman, Margaret Gheen, Avi Kaplan, Revathy Kumar, Michael J. Middleton, Jeanne Nelson, Robert Roeser, and Timothy Urdan. *Manual for the Patterns of Adaptive Learning Scales*. Ann Arbor, MI: University of Michigan, 2000.

Nicholls, John G., P.C. Cheung, J. Lauer, and M. Patashnick. "Individual Differences in Academic Motivation: Perceived Ability, Goals, Beliefs, and Values." *Learning and Individual Differences*, vol. 1, 1989, pp. 63-84.

Ross, Margaret E., Marcy Blackburn, and Sean Forbes. "Reliability Generalization of the Patterns of Adaptive Learning Survey Goal Orientation Scales." *Educational and Psychological Measurement*, vol. 65, no. 3, 2005, pp. 451-464.

Ross, Margaret E., David M. Shannon, Jill D. Salisbury-Glennon, and Anthony Guarino. "The Patterns of Adaptive Learning Survey: A Comparison across Grade Levels." *Educational and Psychological Measurement*, vol. 62, no. 3, 2002, pp. 483-497.

PEABODY PICTURE VOCABULARY TEST, FOURTH EDITION (PPVT-4), 2007

Authors: Lloyd M. Dunn and Douglas M. Dunn		**Type of Assessment:** Individually administered adaptive assessment **Domain:** Language arts/language proficiency (vocabulary)
Publisher: Pearson Assessments 800-627-7271 http://ags.pearsonassessments.com/		**Age/Grade Range:** 2 years, 6 months to 90 years **Administration Interval:** None described
Material, Training, and Scoring Costs: PPVT-4 Test Kit (includes picture easel, manual, 25 performance records, and carrying case): $215 for each form or $390 for both forms PPVT-4 ASSIST scoring: $259		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2 (bachelor's degree with coursework in measurement and domain) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (> 2 hours) Assessors should be thoroughly familiar with the test materials and word pronunciation and should have practiced administering and scoring the assessment.
Languages: English		**Alternate Forms:** Two forms; administration interval not described
Representativeness of Norming Sample: The norming sample consisted of a stratified random sample of 3,540 individuals ages 2 years, 6 months to over 90 years (between 100 and 200 each at one-year intervals for ages 2 to 22 years) selected to match the U.S. population proportionately on gender, race/ethnicity, socioeconomic status, geographic region, and special education status. The sample was restricted to individuals proficient in English. A subsample of 2,003 students in kindergarten through grade 12, based on U.S. Census data on grade distribution, was used to establish grade norms. The assessments were conducted from fall 2005 to spring 2006 at 320 sites nationwide.		**Summary** **Initial Material Cost:** 3 ($200 to $500) **Time to Administer:** 10 to 15 minutes **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 3 (normed within past 10 years and nationally representative)

NARRATIVE

Description: The PPVT-4, an individually administered adaptive assessment designed to measure a student's receptive (auditory) vocabulary level for standard English, is appropriate for people between the ages of 2 years, 6 months and 90 years and above. It has two parallel forms, Forms A and B, each with age-based training practice items and 228 test items grouped into 19 sets of 12 items, with the sets arranged in order of increasing difficulty. During the assessment, the assessor orally presents a stimulus word with a set of four color pictures on an easel and asks the student to identify the picture that best represents the word's meaning. The assessor administers the item sets beginning at a predetermined age-appropriate start item until the basal and ceiling sets are found. On average, students respond to 5 item sets. The basal set is set 1 or the first item set in which the student makes one or no errors. The ceiling set is the first item set in which the student makes eight or more errors or the end of the assessment. Because it requires no reading, writing, or speaking on the part of the student, the PPVT-4 is useful in assessing young students and can be used successfully with individuals with disabilities. Average administration time is 10 to 15 minutes (for 5 sets of 12 items or 60 items).

Other Languages: None.

Uses of Information: The PPVT-4 measures receptive vocabulary in standard English. The publisher reports that the PPVT-4 may also be used to (1) measure an individual's vocabulary growth and/or response to instruction; (2) diagnose reading difficulties; (3) measure language potential, nonreaders' development or impairments, or written- or expressive-language difficulties or other impairments (e.g., aphasia); (4) screen for verbal development; (5) establish rapport with a student as an initial component in a larger battery of assessments; and (6) evaluate the extent of vocabulary of an English learner (though it cannot provide a normative score to use for comparison for such individuals).

Methods of Scoring: The raw score is obtained by subtracting the total number of errors in all sets from the number of the last item in the individual's ceiling set. Raw scores may be converted into age- or grade-normative or developmental scores as well as into a non-normative growth scale value (GSV) score used to measure an individual's improvement over time. Normative scores include standard scores (mean = 100, standard deviation = 15), percentiles, normal curve equivalents (NCE), and stanines. Developmental scores include age and grade equivalent scores. Grade norms are available for kindergarten through grade 12. Using a series of tables, raw scores are converted into GSV, normative, or developmental scores and corresponding confidence intervals. A scoring software program (PPVT-4 ASSIST) that scores, converts scores, and interprets the results is also available for purchase.

Interpretability: Only persons with formal training in psychological testing and statistics should interpret the results of the PPVT-4. The manual provides a brief description of each score as well as uses and limitations. Individuals may compare PPVT-4 scores to previous PPVT administrations by using GSV scores. Qualitative interpretations of incorrect answers may be conducted by using the classification of items by part of speech. The PPVT-4 ASSIST provides score reports, including progress and group reports.

Reliability:
(1) Internal consistency reliability: The Spearman-Brown split-half reliability (within forms) ranged from 0.89 to 0.97 for Form A scores and from 0.91 to 0.97 for Form B scores for those ages 2 years, 6 months to 24 years. Cronbach's alpha for the same age groups ranged from 0.93 to 0.98 for Form A scores and from 0.94 to 0.97 for Form B scores. Calculations of split-half reliabilities were based on separate analysis of the odd and even items in a Rasch analysis. The correlations between forms were adjusted for differences in the standard deviations of the normative sample for each form.
(2) Test-retest reliability: The correlation coefficients ranged from 0.91 to 0.94 between scores from the two administrations (with about a four-week interval) for ages 2 to 14 years (N = 340). No information was provided for individuals ages 15 to 22 years.
(3) Alternate form reliability: The reliability coefficients between Form A and Form B scores (from one session or two sessions up to seven days apart) ranged from 0.83 to 0.90 for students ages 2 to 14 years (N = 508). No information was provided for individuals ages 15 to 22 years.
(4) Inter-rater reliability: No information available.

Validity Evidence:
The pool of stimulus words appropriate for color picture illustration was culled mainly from *Merriam-Webster's Collegiate Dictionary* (2003) and several editions of *Webster's New Collegiate Dictionary* (1953, 1967, 1981) as well as from several other vocabulary or lexicographic resources. Stimulus words were grouped into 20 content categories. The manual details the decisions guiding word selection and picture development for stimulus words and construction of the two parallel forms. The developers paid attention to design of the colorization of the pictures for the PPVT-4 update, including sensitivity to demographic and disability issues. Item difficulty was gauged by using classical and Rasch methods. The authors state that the word stimulus selection process provides qualitative evidence of the content validity of the PPVT-4 as a measure of standard American English receptive vocabulary.

Construct/Concurrent validity: Studies correlated the PPVT-4 with four instruments that measure expressive vocabulary, language ability, and/or reading achievement: the Expressive Vocabulary Test, Second Edition (EVT-2); the Comprehensive Assessment of Spoken Language (CASL); the Clinical Evaluation of Language Fundamentals, Fourth Edition (CELF-4); and the Group Reading Assessment and Diagnostic Evaluation (GRADE). In addition, the PPVT-4 was correlated with the PPVT-III. Assessments were administered on the same day, except for the PPVT-III, which was given up to 11 days later. Sample sizes ranged between 110 and 425 students, except for the EVT-2, which used the same norming sample of 3,540. Students ranged in age from 2 to 24 years but typically were in elementary or middle school. Correlations between the PPVT-4 and the EVT-2 ranged across grades from 0.80 to 0.84. Correlation coefficients with CASL subtest scores ranged from 0.37 to 0.77. Correlations with the CELF-4 language subtest scores ranged from 0.67 to 0.79. Correlations with the GRADE ranged across grades from 0.35 to 0.79 on the total test scores and from 0.27 to 0.79 on vocabulary and comprehension composite scores. Correlations with the PPVT-III scores ranged across grades from 0.79 to 0.83.

Developers examined the difference of PPVT-4 means among nine student groups, including a giftedness group, a language delay and relevant disabilities group, and a non-clinical reference

group from the norming sample (controlling for gender, race/ethnicity, and socioeconomic status). Results showed that all tests were statistically significant.

Predictive validity: Six studies have been conducted with the PPVT-R (the second version of the PPVT) and later achievement, language, and other assessment results. For students in preschool through grade 5, correlations ranged from 0.14 to 0.66.

Bias Analysis: In pre-release trials, the developers conducted item bias analysis by using a Rasch-based method. They eliminated or revised and retested items that, during the first national tryouts, were determined to be biased with regard to gender, race/ethnicity, socioeconomic status, and region of the country; the publisher reports that items determined to be biased during the second national trial were typically dropped from the assessment.

Training Support: Pearson Assessments offers in-service training and content presentations, some in person and some online.

Adaptations/Special Instructions for Individuals with Disabilities: Given that it requires no reading or writing, the PPVT-4 may be administered to many groups with disabilities without any significant changes. The assessor's manual describes various modifications that can be made in administering the assessment to accommodate groups with various disabilities, specifically deaf or hard-of-hearing students. Interpretation of results from the hearing-impaired population should be tentative; an expert on deafness notes that the norms and other standards have not been determined for the hearing-impaired.

Alternate Forms: The PPVT-4 has two parallel forms—Form A and Form B. Administration intervals between forms were not described.

Previous Version: The previous version of the assessment, the PPVT-III, is still used for research purposes. The main updates in the PPVT-4, according to the publisher, include colorized pictures with an increased balance of gender and racial diversity; more stimulus words, particularly at the floor and ceiling of the measure (easiest or most difficult); and GSV scoring, which can be used for measuring a student's progress over time.

NCEE or REL Study Use:[1] The Effects of Opening the World of Learning (OWL) on the Early Literacy Skills of At-Risk Urban Preschool Students**;** Accelerating Language development in kindergarten through Kindergarten PAVEd for Success; Evaluating the Impact of the Program for Infant/Toddler Care; A Study of Classroom Literacy Interventions and Outcomes in Even Start

[1] See Table F.1 for web address.

References:

Dunn, Lloyd M., and Douglas M. Dunn. *Peabody Picture Vocabulary Test–Fourth Edition Manual*. Minneapolis, MN: Wascana Limited Partnership, 2007.

Kisker, Ellen E., Kimberly Boller, Charles Nagatoshi, Christine Sciarrino, Vinita Jethwani, Teresa Zavitsky, Melissa Ford, and John M. Love. "Resources for Measuring Services and Outcomes in Head Start Programs Serving Infants and Toddlers." Washington, DC: Mathematica Policy Research, 2003.

Montgomery, Judy K. *The Bridge of Vocabulary: Evidence-Based Activities for Academic Success.* Minneapolis, MN: Pearson Assessments, 2007.

Pearson Assessments. *ASSIST: Automated System for Scoring and Interpreting Standardized Tests.* Minneapolis, MN: Pearson Assessments, 2006.

Pearson Assessments. "PPVT-4 Publication Summary Form." Available at [http://ags.pearsonassessments.com/pdf/pubsum/ppvt4.pdf]. 2007.

PHONOLOGICAL AWARENESS LITERACY SCREENING (PALS)— PREK, PALS-K, AND PALS 1-3

Authors: Marcia Invernizzi, Anne Sullivan, Joanne Meier, and Linda Swank		**Type of Assessment:** Individual assessment (with some adaptive components) **Domain:** Reading (phonological awareness, letter recognition and naming) and language arts/language proficiency (oral reading, comprehension)
Publisher: University of Virginia, Curry School of Education 1-866-372- PALS http://pals.virginia.edu		**Grade Range/Age:** PALS-PreK: 4-year-olds; PALS-K: kindergarten; PALS 1-3: grades 1 through 3 **Administration Interval:** Fall-spring administrations
Material, Training, and Scoring Costs: PALS-PreK Teacher Set (Administration and Scoring Guide, fall and spring Class Summary Sheets, Student and Teacher Packet, and 20 fall and spring Student Summary Sheets): $75 PALS-PreK Assessment and Training Video: $15 PALS-K and PALS 1-3 Teacher Sets (for 25 students in fall and spring, includes assessment training CD-ROM): $95 each Technical references are available free on developer's web site. The Online Score Entry and Reporting System requires a contract with the PALS office. Cost is not provided.		**Personnel and Training Requirements** **Credentials Required for Use:** No special qualifications required **Personnel for Administration:** Individual with basic clerical skills with some training **Training for Administration:** Minimal (1 to 2 hours) The PALS is available publicly, but the assessor should familiarize him- or herself with the measure by using the training CD-ROM and the Administration and Scoring guides.
Languages: English[1]		**Alternate Forms:** PALS-PreK: none described; PALS-K and PALS 1-3: three forms each
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (under $100) **Time to Administer:** 20 to 25 minutes **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Phonological Awareness Literacy Screening (PALS) assessment consists of three instruments—PALS-PreK (for preschool students), PALS-K (for kindergartners), and PALS 1-3 (for students in grades 1 through 3)—that measure reading and language skills as well as knowledge of the English writing system. Each instrument is individually administered with some elements that are adaptive such that additional subtests (called tasks by the developer) may be administered depending on performance on another subtest as detailed below. The administration time is 20 to 25 minutes for any one of the three instruments.

The PALS-PreK measures early phonological and print awareness in 4-year-olds and consists of a total 99 items across six subtests: Name Writing, Alphabet Knowledge, Beginning Sound Awareness, Print and Word Awareness, Rhyme Awareness, and Knowledge of Nursery Rhymes. The Alphabet Knowledge subtest consists of three components: Upper Case Alphabet Recognition, Lower Case Alphabet Recognition, and Letter Sounds. The latter two components are adaptive (that is, the student must score at a predetermined level or above on the previous component before administration of subsequent components).

PALS-K measures knowledge of speech sounds and awareness of print in kindergarten students. It consists of 114 items across six required subtests: Group Rhyme Awareness, Group Beginning Sound, Alphabet Recognition, Letter Sounds, Spelling, and Concept of Word. The Group subtests and Spelling subtest may be administered first in small groups; the remaining three subtests (Alphabet Recognition, Letter Sounds, and Concept of Word) are then administered individually. If the student does not reach a particular benchmark for the Group subtests, the assessor administers additional individual subtests: Individual Rhyme Awareness and Individual Beginning Sound, respectively. Hence, the PALS-K may be considered adaptive for the areas of rhyme awareness and beginning sound. An optional subtest for Word Recognition in Isolation is available with three components: Preprimer List, Primer List, and First Grade List. The authors do not note administration time for these additional components.

PALS 1-3 is an individual adaptive assessment in which students complete additional subtests if they do not provide enough correct responses to reach a predetermined benchmark as a condition of a more in-depth assessment of oral reading, alphabetic knowledge, comprehension, and speech sounds. The assessment consists of four levels, in which the first is Entry Level with two subtests (Word Recognition and Spelling). If students meet the benchmarks at this level, they stop; otherwise, they complete Level A with four subtests (Oral Reading Accuracy, Oral Reading Fluency, Oral Reading Rate, and Oral Reading Comprehension). Again, if they fail to meet the benchmarks, they are routed to Level B with three subtests (Alphabet Recognition, Letter Sounds, and Concept of Word) and then finally, if necessary, to Level C with two subtests (Blending and Sound-to-Letter). The authors note that the Word Recognition subtest at the Entry Level has a highly restricted score range and is thus reported to have a ceiling effect, with most students' scores clustered at the top of the scale (that is, most students meet the competency level).

Other Languages: A Spanish version of the PALS is currently being piloted, but no further information is yet available on it.

Uses of Information: The PALS measure students' reading and language development and knowledge of the English writing system and may assess growth and monitor progress. The PALS-K and PALS 1-3 also function as screening tools to identify students who are performing below grade level for literacy, thus helping to pinpoint students possibly at risk for reading difficulties and delays. The PALS-PreK is not considered a screener; instead, its authors describe it as a diagnostic tool to assess current knowledge of the English language and writing system.

Methods of Scoring: For all three instruments, assessors calculate raw scores for each student on each of the subtests by using a Child Summary Sheet (which consists of the actual items and a space to record the number of correct responses under each subtest). For the PALS-K and the PALS 1-3, assessors also enter the summed score on the Child Summary Sheet. Assessors use the Class Summary Sheet to enter subtest scores and summed scores (if applicable) for several students or the entire class. The PALS web site provides a tool (Online Score Entry and Reporting System) that allows the assessor to enter raw scores to obtain reports for groups of students (for example, for a class). The system includes an Online Assessment Wizard that guides the assessor in entering scores online while screening students and provides instructional reports for teachers and specialists. To use the Online Score Entry and Reporting System, the assessor must enter into a contract with the PALS office. Additional information on scoring and administration procedures may be found in the Administration and Scoring guides for the PALS-K and PALS 1-3 (PALS-PreK does not include a separate Administration and Scoring Guide).

Interpretability: The raw scores are compared to developmental ranges or benchmarks, which provide an indication of the minimal level of competency for a given subtest. For PALS-PreK, developmental ranges were based on field testing, several pilot studies (described in the Validity Evidence section), longitudinal analyses data using PALS-K and PALS 1-3, and PALS-K benchmarks. For PALS-K and PALS 1-3, developers undertook a benchmarking process to specify cut scores. The developers indicated that classroom teachers commonly interpret and apply the results.

Reliability:
(1) Internal consistency reliability: The information below is based on a variety of investigations that estimated the reliability of scores conducted from 1998 to 2005 by using data from pilot studies and statewide data samples. For PALS-PreK scores, based on samples ranging from 99 to 138 students, Cronbach's alpha was reported for four of six subtests and ranged from 0.75 (Print and Word Awareness) to 0.93 (Beginning Sound). Alphas associated with Name Writing and Alphabet Knowledge scores were not reported as the subtests used the child's own name or included all letters, respectively. Guttman split-half reliability coefficients were also reported for the scores from the same four subtests, ranging from 0.71 (Print and Word Awareness) to 0.94 (Beginning Sound). For PALS-K, Cronbach's alphas for the scores on all subtests ranged from 0.79 to 0.89 across socioeconomic status and gender groups. For PALS 1-3, internal consistency was examined for the scores from a grade 1 cohort across two separate years, with alphas ranging from 0.66 to 0.88 for summed scores across gender, socioeconomic groups, geographic region, and ethnicities. Alphas for the total sample ranged from 0.76 to 0.83 across the two years for the summed scores.
(2) Test-retest reliability: No information was provided on test-retest reliability for scores from the PALS-PreK. For PALS-K scores, test-retest reliability estimates, based on a one- to two-week interval between administrations, ranged from 0.78 to 0.95 for a sample of 473

kindergarten students. For PALS 1-3 scores, test-retest reliability estimates, based on a one- to two-week interval between administrations, ranged from 0.88 to 0.97 for the Entry Level subtest scores, with a sample of 204 grade 1, 2, and 3 students.
(3) Alternate form reliability: PALS-PreK has no alternate forms. For PALS-K and PALS 1-3, no alternate form reliability evidence was described.
(4) Inter-rater reliability: Scoring consistency was measured when one person administered the assessment while a second person observed and scored subtest items simultaneously but independently. For PALS-PreK, the correlations between scorers were 0.99 for all subtests (based on samples of 99 to 138 students), except for Print and Word Awareness, which was not reported. For PALS-K, inter-rater reliability between scorers ranged from 0.96 to 0.99 across subtests, based on samples of 121 to 154 students. For PALS 1-3, inter-rater reliability was based on five years of data on samples ranging from 18 to 375 students. Correlations between scorers ranged from 0.81 to 0.99 for Entry Level and Level B and C subtests. Correlations between scorers ranged from 0.63 to 0.97 for the Level A subtests.

Validity Evidence:
An advisory panel composed of experts in early literacy development ensured the assessment of items representing key subject matter. For each subtest, developers determined the best representation of items in a variety of ways. The PALS-PreK was pilot tested on preschool students in Virginia during five studies conducted from 2000 to 2005. In an early study, developers conducted factor analysis for a preschool sample of 56 students. Results showed that PALS-PreK measures one trait, namely, emergent literacy. The PALS-K and PALS 1-3 were studied, refined, and validated by using statewide data collected since fall 1997 and data from three pilot studies conducted from spring 2001 through spring 2004. The PALS 1-3 was also studied in a small initial pilot in spring 2000. Samples in the PALS-K statewide cohorts ranged from 37,072 to 83,934 students and 1,772 to 3,924 students in the pilot studies. Samples in the PALS 1-3 statewide cohorts ranged from 140,000 to 150,000 students and 13,021 grade 1, 2, and 3 students in the pilot studies.

Construct/Concurrent validity: Developers conducted principal components analysis (PCA) by using PALS-K and PALS 1-3 data each year since 1997. The PCA results indicated that PALS-K and PALS 1-3 each assesses a single construct—beginning reading. For Nursery Rhyme Awareness, alpha coefficients were examined for 30 nursery rhymes piloted to select the 10 highest coefficients, indicating an absence of floor effects. A similar process was used for the PALS-PreK, PALS-K, and PALS 1-3.

With respect to convergent validity, PALS-PreK was compared to three existing measures: Sawyer's Test of Awareness of Language Segments (TALS) Part A, High/Scope's Child Observation Record (COR), and the Test of Early Reading Ability (TERA-3). The studies are based on different pilots, using earlier as well as revised versions of the PALS-PreK. The correlations were 0.41 between the PALS-PreK and the TALS Part A scores, 0.67 between the PALS-PreK scores and the Alphabet, Conventions, and Meaning subtest scores on the TERA-3, and 0.71 between the PALS-PreK summed score and the COR language and literacy component scores. Sample sizes for each comparison ranged from 70 to 90 students. Authors do not specify which PALS-PreK subtests were used in the comparisons with the TALS Part A or the TERA-3.

The PALS-K summed score was correlated with the Stanford Achievement Test, Ninth Edition (SAT-9) Total Reading scaled score and three Reading subtests (Sounds and Letters, Word Reading, and Sentence Reading) on a sample of 137 kindergartners. The same students had been administered PALS-K two weeks earlier. The correlation between the PALS-K summed score and the SAT-9 Total Reading scaled score was 0.72, ranging from 0.58 (Sentence Reading) to 0.79 (Sounds and Letters) across subtests.

For the PALS 1-3, developers correlated the Entry Level summed score with five measures based on samples of 200 to 300 students in grades 1, 2, and 3. Correlations with reading assessments (for example, the Qualitative Reading Inventory-II, Developmental Reading Assessment, California Achievement Test [CAT/5], and SAT-9 Reading scale) ranged from 0.57 to 0.81. The PALS 1-3 was correlated with the Virginia Standards of Learning (SOL) Total Reading at 0.57. The Entry Level Spelling subtest was also examined in relation to the word analysis subtests of the SOL and CAT/5, correlating at 0.52 and 0.66, respectively, and with the CAT/5 Total Reading score at 0.70.

For PALS-K, authors used discriminant analysis to report that 98 percent of students were classified accurately as identified for reading needs on all subtests (consistent with data collected annually since 1997). For PALS 1-3, discriminant analysis functions conducted for the Entry Level subtests accurately classified 93 to 98 percent of students across grades 1, 2, and 3.

Predictive validity: Developers compared the PALS-PreK with subsequent results from PALS-K and PALS 1-3 to determine predictive validity. With separate samples, the PALS-PreK correlated with the PALS-K in the fall at 0.91, with the PALS-K in the spring at 0.53, and with the PALS 1-3 in the fall of grade 1 at 0.56. The first correlation was obtained with a sample of 41 students while the latter two involved over 2,500 students each. Discriminant function analysis showed that scores on the PALS-PreK subtests predicted the correct classification for 86.5 percent of students as needing additional reading instruction as determined by the PALS-K the following fall and the correct classification for 73.5 percent of students as determined by the PALS 1-3 two years later.

Using a sample of 74 kindergartners ($r = 0.70$), developers correlated the PALS-K fall scores with the spring SAT-9 scores. For the same sample of kindergartners, the fall PALS-K scores correlated with the spring PALS-K scores at 0.56 and later with PALS 1-3 scores at 0.67 and 0.53 in the fall and spring, respectively. Discriminant analysis for a sample of 799 students was conducted to assess the relationship between the SOL reading scores from spring of grade 3 and (1) PALS-K scores from the same year, (2) PALS scores from fall of grade 2, and (3) PALS scores from spring of grade 3. Developers indicate that students' combined PALS scores predicted the correct classification for 82 percent as passing or failing the SOL in grade 3.

Developers assessed predictive validity of fall PALS 1-3 scores by correlating them with (1) spring SAT-9 scores for over 700 grade 1 and 2 students and (2) Virginia's spring SOL reading assessment scores for 277 grade 3 students. Correlations were 0.73 and 0.63 for grade 1 and 2 students between the PALS 1-3 Entry Level summed score and the SAT-9 Total Reading Scaled Score and 0.60 when correlated with the SOL Total Reading Score. Developers also correlated PALS 1-3 spring scores of separate samples of grade 1 and 2 students to PALS 1-3 fall scores of the same samples (in grades 2 and 3). Entry Level Word Recognition and Spelling subtests

indicated correlations of 0.79 and 0.81 for grade 1 students entering grade 2 and 0.81 and 0.83 for grade 2 students entering grade 3.

Bias Analysis: Developers convened two advisory review panels comprised of preschool teachers, early childhood program coordinators, faculty members, and other educators to examine the content of the PALS and individual items for difficulty, bias, clarity, and consistency. Panel members provided comments to PALS staff as part of the measure's development and field testing.

Training Support: The Teacher's Manual recommends that users watch the assessment training video as well as read and understand the Teacher's Manual and, in the case of PALS-K and PALS 1-3, the Administration and Scoring Guide. Training videos and CDs are available for purchase through the developer's web site. The web site also offers tools to help with test administration and scoring by means of the PALS Online Score Entry and Reporting System and provides online demonstrations for administering the assessments.

Adaptations/Special Instructions for Individuals with Disabilities: Large print editions, Braille editions, and closed-captioned video are available through Barbara Jones at the Virginia Department of Education, Office of Elementary Instructional Services (804-786-1997; bjones@mail.vak12ed.edu).

Alternate Forms: The PALS-PreK does not have alternate forms. Three forms are available for PALS-K and for PALS 1-3: Forms A, B, and C. Previous use included the same form (A or B) within a given year as frequently as desired and then an alternate form the following year. The PALS-K Form C was used previously as a mid-year form in any given year.

Previous Version: None.

NCEE or REL Study Use:[2] The Effects of Opening the World of Learning (OWL) on the early Literacy Skills of At-Risk Urban Preschool Students

[1] Spanish version is being piloted currently, but no information is available on the web site.

[2] See Table F.1 for web address.

References:

Invernizzi, Marcia, Joanne Meier, and Connie Juel. *PALS 1-3 Administration and Scoring Guide Form A*. Richmond, VA: The Rector and The Board of Visitors of the University of Virginia, 2003–2005.

Invernizzi, Marcia, Joanne Meier, Linda Swank, and Connie Juel. *PALS-K Administration and Scoring Guide Form A*. Richmond, VA: The Rector and The Board of Visitors of the University of Virginia, 2003–2005.

Invernizzi, Marcia, Joanne Meier, Linda Swank, and Connie Juel. *PALS-K Administration and Scoring Guide Form B*. Richmond, VA: The Rector and The Board of Visitors of the University of Virginia, 2004.

Invernizzi, Marcia, Amie Sullivan, Joanne Meier, and Linda Swank. *PALS-K Technical Reference Form A*. Richmond, VA: The Rector and The Board of Visitors of the University of Virginia, 2003–2007, Available at [http://pals.virginia.edu/pdfs/rd/tech/K_Tech_Ref_07-08.pdf].

Invernizzi, Marcia, Amie Sullivan, Joanne Meier, and Linda Swank. *PALS-PreK Teachers Manual*. Richmond, VA: University of Virginia, 2004.

Invernizzi, Marcia, Amie Sullivan, Joanne Meier, and Linda Swank. *PALS 1-3 Technical Reference*. Richmond, VA: The Rector and The Board of Visitors of the University of Virginia, 2003–2007, Available at [http://pals.virginia.edu/pdfs/rd/tech/1-3_TechRef_07-08_FormA.pdf].

PRELAS 2000, 1998

Authors: Sharon E. Duncan and Edward A. De Avila		**Type of Assessment:** Individual assessment **Domain:** Reading (letter, number, and word recognition), language arts/language proficiency (expressive and receptive oral language skills, grammar, syntax, vocabulary, morphology, writing), mathematics (number concepts), and basic concepts (colors, shapes, and spatial relationships)
Publisher: CTB/McGraw-Hill 800-538-9547 http://www.ctb.com/		**Grade/Age Range:** Prekindergarten through grade 1, age 4 through 6 years **Administration Interval:** None described
Material, Training, and Scoring Costs: English Examiner's Kit (examiner's manual, quick reference guide, cue picture book, game board, audio cassette, 50 answer sheets): $239.50 LASscore Basic Module (optional): $391.75		**Personnel and Training Requirements** **Credentials Required for Use:** Level 1 (training or supervised experience with measurement) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours) Assessors must be proficient English speakers, qualified to work with 4- through 6-year-olds, and familiar with all aspects of test administration.
Languages: English and Spanish		**Alternate Forms:** Two forms; administration interval not described
Representativeness of Norming Sample: The norms were based on 965 4- through 6-year-olds sampled in 1997. The convenience sample was drawn from four regions, with half of the sample from Midwestern cities. The students' age and grade were noted; 88 students were 4-year-olds and 877 students were older than 4 years, and 11 percent were in prekindergarten, 54 percent in kindergarten, and 35 percent in grade 1. Most students spoke a language other than English at home, predominantly Spanish, including 50 percent non–English speakers and 23 percent English and other-language speakers. Students were drawn from moderate- to low-income households. English-only speakers demonstrated average scores on a standardized achievement test.		**Summary** **Initial Material Cost:** 3 ($200 to $500) **Time to Administer:** 15 to 25 minutes **Ease of Administration and Scoring:** 3 (administered and scored by highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 2 (older than 10 years or not nationally representative)

NARRATIVE

Description: The English PreLAS 2000 is an individual assessment that measures English expressive and receptive language and preliteracy skills. The assessment is normed for 4- through 6-year-olds. The assessment consists of two tests: Oral Language and Preliteracy. The average administration time for the PreLAS 2000 is 15 to 25 minutes, including 10 to 15 minutes for the Oral Language test and 5 to 10 minutes for the Preliteracy test. The Oral Language test consists of five subtests with 40 discrete items (i.e., assessing an isolated element of language) and 2 short story items (i.e., assessing several elements of language related to English proficiency). The subtests include Simon Says, Art Show, Human Body, Say What You Hear, and Let's Tell Stories. The Simon Says subtest assesses receptive language by using a game format in which students respond to simple directives with physical responses. The Art Show and Human Body subtests involve the labeling of items displayed in a picture book and require students to name items and their purpose. The Say What You Hear subtest assesses receptive and expressive abilities; students repeat sentences focusing on particular grammatical forms such as negative commands and imperatives, plural words, and contractions. The Let's Tell Stories subtest assesses expressive language, requiring students to listen to either audio-cassette recordings or the assessor's recitation of short narrations and then watch as assessors point to supporting pictures displayed in a picture book. Students then retell the narration in their own words, using pictorial support. Preliteracy is administered to 5- and 6-year-olds only and includes 30 items. The Preliteracy test includes six subtests: Letters, Numbers, Colors, Shapes and Spatial Relationships, Reading, and Writing. A game board is used for the first five subtests. For the Writing subtest, the assessor dictates what the student should write (i.e., name, age, and two- and three-letter words).

Other Languages: The Spanish PreLAS 2000 was normed on 397 native Spanish-speaking students. The convenience sample was drawn from 11 sites in the United States (38 percent of students) and Latin America (62 percent). The sample consisted of 208 students age 4 years and 189 students older than 4 years. U.S. students were younger on average than Latin American students (60 and 66 months, respectively). The sample selection process for the Spanish version differed from that for the English version in that developers could not directly match socioeconomic and cultural factors between students from the United States and students from Latin American countries. Spanish and English items on Form C are identical in structure, format, methods, and techniques but differ in terms of content, words, syntax, and artwork. No comparability tests were reported between the Spanish and English PreLAS 2000.

Developers provided separate internal consistency and discriminant analysis data for the Spanish version (De Avila and Duncan 2000). Cronbach's alphas were calculated for each test and subtest (excluding Let's Tell Stories). For the Oral Language test, coefficients for subtests ranged from 0.64 to 0.81. For the Preliteracy test, coefficients ranged from 0.55 to 0.96.

Authors compared mean test scores by country of origin and age using *t*-test group comparisons; sample sizes ranged from 228 to 397. Students from Latin America had higher scores than students from the United States on all subtests within the Oral Language and Preliteracy tests. Older students had higher scores on some subtests, including Let's Tell Stories (Oral Language) and Letters and Reading (Preliteracy). Developers noted that schooling differences between

different Latin American sites might be attributable to inconsistent test score differences for older students.

Uses of Information: The English PreLAS 2000 assesses language proficiency in young students from homes in which English is not the first language. The Spanish PreLAS 2000 assesses constructs identical to those in the English PreLAS 2000 to determine Spanish language proficiency. Developers define language proficiency as linguistic elements necessary for successful communication within the school environment; proficiency involves listening, speaking, reading, and writing. Developers assert that language proficiency may be used to help define academic achievement.

Sponsored by the National Center for Education Statistics, the Early Childhood Longitudinal Study–Kindergarten Class of 1998–1999 (ECLS–K) adopted the English and Spanish language versions of three Oral Language subtests (Simon Says, Art Show, and Let's Tell Stories) and distinguished this adapted measure as the Oral Language Development Scale (OLDS) (Rock and Pollock 2002). The English OLDS was used as an English-language screener for the study's cognitive assessment battery. To measure Spanish knowledge, the Spanish OLDS was administered to students who did not pass the English OLDS.

Methods of Scoring: Raw scores are calculated by adding the total points obtained for each of the five subtests in the Oral Language test and each of the six subtests in the Preliteracy test. Raw scores are converted to weighted scores for each subtest, which are summed for a total weighted test score. A table is available to convert weighted scores to proficiency levels. Each test has one holistically scored subtest: Let's Tell Stories (Oral Language) and Writing (Preliteracy). For Let's Tell Stories, assessors judge student story retelling based on six categories of criteria describing language performance (for example, no response in English would be 0, and fluent English response with vivid vocabulary and complex constructions would be 5). Assessors may calculate scores for individuals and groups manually or use the publisher's computer software. Graphics of assessment results may be created with computer software.

Interpretability: A table in the manual provides details on how to interpret assessment results. The Oral Language test weighted score is converted to one of five proficiency levels (from 1 = non–English speaker to 5 = fluent English speaker), by age group (4-year-olds and 5- and 6-year-olds). The Preliteracy test weighted score is converted to one of three proficiency levels (low, mid-level, high) for 5- and 6-year-olds only. An appendix in the manual describes score ranges in which student proficiency levels may be misclassified and notes the steps to take when students have scores in those ranges.

Reliability:
(1) Internal consistency reliability: Cronbach's alphas were calculated for each subtest (excluding Let's Tell Stories), separately by form. For the Oral Language test, coefficients for Forms C and D subtest scores both ranged from 0.86 to 0.90. For the Preliteracy test, coefficients for Form C subtest scores ranged from 0.84 to 0.91, and coefficients for Form D subtest scores ranged from 0.76 to 0.92. The ECLS–K (Rock and Pollock 2002) reported split-half correlation coefficients for the English OLDS, which was completed only by students who spoke a language other than English; the coefficients ranged from 0.96 to 0.98 across kindergarten and grade 1. The sample decreased over time (2,865 to 945) because students were exempted if they passed

the English OLDS in earlier waves. On the Spanish OLDS, coefficients ranged from 0.91 to 0.92, and the tested sample decreased from 1,039 to 370. The ECLS–K psychometric report shows that the individual OLDS subtests, though varying in weight toward the total score, all ranged from the mid-0.80s to the mid-0.90s (Rock and Pollack 2002).
(2) Test-retest reliability: Correlation coefficients were calculated by subtest within each PreLAS 2000 test. For the Oral Language test, subtest correlation coefficients between scores from two administrations ranged from 0.76 (Let's Tell Stories) to 0.94 (Art Show). For the Preliteracy test, subtest correlation coefficients between scores from two administrations ranged from 0.79 (Reading) to 0.96 (Writing). The interval between administration sessions was not described.
(3) Alternate form reliability: Two English PreLAS 2000 forms are available, C and D. Correlation coefficients between forms for the Oral Language and Preliteracy tests were 0.99 and 0.97, respectively. Subtest coefficients between scores on the two forms in the Oral Language test (excluding Let's Tell Stories) ranged from 0.87 (Simon Says) to 0.99 (Human Body). Subtest coefficients between scores on the two forms in the Preliteracy test ranged from 0.79 (Reading) to 0.96 (Writing). In the Let's Tell Stories subtest, scores for three stories in Form C were compared to scores for three stories in Form D. Coefficients ranged from 0.76 to 0.96 for two Form C stories correlated with the Form D stories. One Form C story scores, Butterfly, had lower coefficients when compared to scores on the Form D stories, ranging from 0.59 to 0.83.
(4) Inter-rater reliability: Correlation coefficients were computed on "interjudge" ratings for two subjectively scored subtests. The procedures were not described. In the Oral Language test, the Let's Tell Stories subtest coefficient was 0.88. In the Preliteracy test, the Writing subtest coefficient was 0.90.

Validity Evidence:
Professional staff, mostly teachers, wrote the PreLAS 2000 items and stories after researching the literature on child language and kindergarten readiness skills, bilingualism, assessment of immigrant children, conducting observations of preschools and kindergarten instruction, and soliciting expert opinion. Items were matched to a theoretical rationale, carefully reviewed for content and accuracy, and pilot tested and analyzed. An expert panel reviewed artwork for the final selection of items. Developers selected phonologically and intellectually appropriate vocabulary and compared the Let's Tell Stories passages to the 1985 PreLAS stories for comparable story length, sentence length, mean number of words and sentences, and communication unit analyses.

Construct/Concurrent validity: No information was provided on construct or concurrent validity against other measures. Developers, however, compared mean test scores by age, grade, language group, and interactions between these groups by using ANOVA techniques; sample sizes ranged from 880 to 961. Developers presented Form C results only, noting that Form D results did not differ considerably. Oral Language and Preliteracy test mean scores increased significantly with age and by grade. English-only speakers had significantly higher Oral Language and Preliteracy test total scores and individual subtest scores than students who spoke a language other than English. Mean test scores for other language speakers varied significantly more than for English-only speakers. When age and language interactions were examined on the Oral Language test, English-only speakers had consistently higher mean scores at both age levels, whereas non-English speakers had greater score increases with the increase in age. Interactions between grade and language factors yielded similar results on the Oral Language test in that differences in scores between English-only and non-English speakers became smaller as

grade level increased. Similar trends, although non-significant, were found for the Preliteracy test when age-language and grade-language interactions were examined. Score cutoffs introduce misclassification of proficiency levels in two instances. The probability of Oral Language proficiency level misclassification is higher when the student falls in the upper range of the third proficiency level (limited English speaker) and the lower range of the fourth proficiency level (fluent English speaker). The probability of Preliteracy proficiency-level misclassification is higher when students fall in the upper range of the mid-level proficiency level and the lower range of the high-level proficiency level.

Predictive validity: No information available.

Bias Analysis: Oral Language test mean scores on Form C did not differ significantly by gender. Mean scores for male and female students were not presented for the Preliteracy test.

Training Support: The manual guides assessors on planning the testing schedule, preparing the test environment and the students, following standardized test procedures, administering the test, and completing answer sheets. For the Let's Tell Stories subtest, the manual provides assessors with a separate holistic scoring chapter. For this subtest, two assessors (or two teams of assessors) separately collect and score at least 10 responses from students in one age group. The first assessor's/team's score must agree with the second assessor's/team's score on at least 9 out of 10 stories. If there is a score disagreement on one story, it may not exceed one rating point. The publisher provides software training for purchasers of computerized scoring software.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: Two English forms are available, including (1) Form C Initial Placement and Identification of Students and (2) Form D Student Progress and Redesignation. The recommended time between administrations is not described. Form C is available in Spanish.

Previous Version: The PreLAS 2000 is a revision of the 1985 PreLAS forms (A and B) and is a lower extension of the Language Assessment Scales (LAS) developed by the same authors for students and adults (Pratt 2003).

NCEE or REL Study Use:[1] National Evaluation of Early Reading First

[1] See Table F.1 for web address.

References:

CTB/McGraw-Hill. *LASscore K-12 Basic Module*. Monterey, CA: CTB/McGraw-Hill.

De Avila, Edward A., and Sharon E. Duncan. *PreLAS 2000 English and Spanish Technical Notes*. Monterey, CA: CTB/McGraw-Hill, 2000.

Duncan, Sharon E., and De Avila, Edward A. *PreLAS 2000 Examiner's Manual English Forms C and D*. Monterey, CA: CTB/McGraw-Hill, 1998.

Pratt, Sheila. "Review of the PreLAS 2000." In *The Fifteenth Mental Measurements Yearbook*, edited by Barbara S. Plake, James C. Impara, and Robert A. Spies. Lincoln, NE: The Buros Institute of Mental Measurements, 2003.

Rock, Donald A., and Judith M. Pollock. "Early Childhood Longitudinal Study--Kindergarten Class of 1998-99 (ECLS-K), Psychometric Report for Kindergarten through First Grade." (NCES 2002-05). Washington, DC: National Center for Education Statistics, U.S. Department of Education, 2002.

PRESCHOOL INDIVIDUAL GROWTH AND DEVELOPMENT INDICATORS (IGDI), 1998

Authors: Scott McConnell		**Type of Assessment:** Individual assessment **Domain:** Reading (phonological awareness) and language arts/language proficiency (expressive language)
Publisher: Center for Early Education and Development College of Education and Human Development University of Minnesota 612-625-3058 http://www.cehd.umn.edu/ceed/		**Grade/Age Range:** Preschool students age 3 through 5 years **Administration Interval:** As frequently as desired, but probably no more than monthly
Material, Training, and Scoring Costs: Stimulus cards, record forms, administration instructions, and an administration checklist are available for each IGDI for download from the web site. The web site also provides tools for managing student data and generating reports.		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2 (bachelor's degree with coursework in measurement and domain) **Personnel for Administration:** Individual with basic clerical skills with some training **Training for Administration:** Minimal (1 to 2 hours) The developer suggests that experienced assessors monitor assessments conducted by individuals with minimal background in test administration. They also recommend that assessors become thoroughly familiar with the materials and practice administering the assessment before official use.
Languages: English		**Alternate Forms:** Yes with random selection of items each administration, as frequently as desired, but no more than monthly
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (under $100) **Time to Administer:** 1 to 2 minutes per IGDI **Ease of Administration and Scoring:** 2 (self-administered or administered and scored by someone with basic clerical skills) **Reliability:** 2 (all or mostly under 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description:[1] The Preschool Individual Growth and Development Indicators (IGDI) are a group of timed, individually administered measures that assess the development of early literacy and language skills in preschool students.[2] The following three IGDIs are available from the Get it Got it Go! web site administered by the University of Minnesota: (1) Picture Naming, (2) Rhyming, and (3) Alliteration. Each assessment involves the use of a randomly selected series of picture cards presented to the student. For the Picture Naming IGDI, the student has one minute to name as many pictures as possible from a set of approximately 100 cards. For the Rhyming and Alliteration IGDIs, the assessor points to and names the picture on the top of the card and then the three pictures below. For Rhyming, the assessor then asks the student to point to the picture in the bottom row that sounds the same as the top picture. For Alliteration, the assessor asks the student to point to the picture that starts with the same sound as the top picture. In each of the two IGDI, the student has two minutes to respond to as many pictures as possible from a set of approximately 50 cards. The picture cards, administration instructions, a checklist, and recording forms for each assessment are available for download from the web site. For the Rhyming and Alliteration IGDIs, the assessor must know the correct answers before administering the assessment because the picture cards do not indicate the correct response. The assessments are progressively more difficult and should be administered in the following order: Picture Naming, Rhyming, and Alliteration. If a student does not correctly respond to any of the Picture Naming cards, the assessor should not administer Rhyming or Alliteration.

Other Languages: The Picture Naming IGDI has been translated into Spanish. No psychometric information is available for the Spanish version.

Uses of Information: The preschool IGDIs are designed to monitor students' early literacy and language development through the use of measures assessing their expressive language and phonological awareness skills. The assessments may be used to screen students and identify instructional needs. Repeated administrations of the IGDIs provide a picture of a student's development over time and may assess intervention efforts. The developer emphasizes that IGDIs are not intended to be used as diagnostic tools and notes that additional information and assessments should guide instructional planning when a student's results indicate the need for intervention.

Methods of Scoring: A student's total score is the number of cards answered correctly within the specified time period. Under the supervision of an experienced assessor, individuals with basic clerical skills and some training may administer and score the IGDIs.

Interpretability: The developer recommends the interpretation of a student's performance to include the student's current score as well as his or her rate of growth over at least three assessments. After entering student information on the Get it Got it Go! web site, the assessor may generate reports to compare a student's performance to a criterion group based on a study of typically developing English-speaking preschool students.

Reliability:
(1) Internal consistency reliability: Given the random selection of cards, no two administrations are exactly the same (see Alternate form reliability for information on reliability).
(2) Test-retest reliability: Based on a sample of 29 preschool students over a three-week test interval, the reliability coefficient for Picture Naming scores was 0.67 (Missall et al. 2006). The reliability coefficient over a three-week test interval with a sample of 42 preschool students ranged from 0.83 to 0.89 for Rhyming scores and from 0.62 to 0.88 for Alliteration scores (Missall et al. 2006).
(3) Alternate form reliability: Based on a one-month test interval, the alternate form reliability for Picture Naming ranged from 0.44 to 0.78 (McConnell et al. 2002).
(4) Inter-rater reliability: No information available.

Validity Evidence:
The research team conducted literature reviews of child development outcomes and child assessments, resulting in the identification of 15 socially valued child development outcomes for children from birth to 8 years of age. The team then surveyed more than 1,000 parents and early childhood education specialists, asking them to rate the importance of these outcomes. The most highly rated was "Child uses gestures, sounds, words or sentences to convey wants and needs or to express meaning to others" (Missall et al. 2008; McConnell et al. 1998). Additional literature reviews determined which skills are necessary to achieve the specific outcomes. For example, for the language outcome listed above, one of the precursor skills identified in the literature review was the ability to produce discrete words. The IGDIs were based on the concrete skill sets identified by the literature review (Missall et al. 2008). The team initially developed 10 preschool measures, but they only disseminated Picture Naming, Rhyming, and Alliteration because the other assessments lacked adequate psychometric properties (Missall and McConnell 2004).

Construct/Concurrent validity: A sample of 90 typically developing students and students with disabilities age 36 to 60 months were administered the Picture Naming IGDI, the Peabody Picture Vocabulary Test, Third Edition (PPVT-3), and the Preschool Language Scale-3 (PLS-3). Correlations between scores on Picture Naming and the PPVT-3 ranged from 0.56 to 0.75 while correlations between scores on the Picture Naming and the PLS-3 ranged from 0.63 to 0.79 (Missall et al. 2006). With children age 24 to 44 months, the correlation between scores on Picture Naming and the PLS-3 ranged from 0.74 to 0.81 (Missall and McConnell 2004). The Picture Naming IGDI and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Letter Naming Fluency (LNF) and Onset Recognition Fluency (ORF) subtests were administered to 154 students age 48 to 83 months. The correlations between Picture Naming scores and scores on the LNF ranged from 0.26 to 0.37 across a fall and winter assessment. Correlations between Picture Naming and ORF scores ranged from 0.32 to 0.49 (Missall 2002).

Based on an analysis with 90 students (including students with disabilities), correlations between the Rhyming IGDI scores and scores on related measures of phonological awareness ranged from 0.56 to 0.62 for the PPVT-3, from 0.54 to 0.64 for Concepts about Print (CAP), and from 0.44 to 0.62 for the Test of Phonological Awareness (TOPA) (Missall et al. 2006). Based on a sample of 154 students age 48 to 83 months, correlations between the Rhyming IGDI scores and scores on the DIBELS LNF subtest ranged from 0.48 to 0.58 across a fall and winter assessment

and from 0.44 to 0.68 between Rhyming scores and scores on the DIBELS ORF subtest (Missall 2002).

Based on a sample of 90 students (including students with disabilities), correlations between the Alliteration IGDI scores and scores on related measures of phonological awareness ranged from 0.40 to 0.57 for the PPVT-3; from 0.75 to 0.79 for the TOPA; and from 0.34 to 0.55 for the CAP (Missall et al. 2006). Correlations between Alliteration IGDI scores and scores on the DIBELS LNF subtest ranged from 0.39 to 0.71 (Missall et al. 2006).

Developers also compared assessment scores for several subgroups. Overall correlations between students' age and their Picture Naming score ranged from 0.41 in a longitudinal study with 90 students to 0.60 in a cross-sectional study with 39 students. The correlation between scores and age was 0.63 for typically developing students, 0.32 for students in Head Start, and 0.48 for students with disabilities (McConnell et al. 2002). Based on a sample of 58 preschool students with and without disabilities, the correlation between students' age and their scores was 0.46 for the Rhyming IGDI and 0.61 for the Alliteration IGDI (Missall et al. 2006). The developers conducted hierarchical linear modeling (HLM) growth curves for all three IGDI measures with a sample of 90 students, estimating status at a particular age and rate of growth per month for typically developing students, low-income students, and students with disabilities. They did not, however, note the significance of the coefficients for status and growth (Missall et al. 2006).

A separate HLM analysis with data from 69 students showed a significant difference in the rate of growth across the full sample for Rhyming scores but no significant differences for Picture Naming or Alliteration scores. Across IGDIs, approximately 73 percent of the variance in scores was attributable to students' maturation over a five-month span. Using the same sample of 69 students, the developers conducted a second HLM analysis that contrasted a typically developing control group against Head Start students, students with disabilities, and Spanish-speaking students learning English (or English Language Learners [ELL]). For Picture Naming, the scores of the ELL students were significantly lower than those of the control group. For Rhyming and Alliteration, the scores of all the groups were significantly lower than those of the control group. The rate of growth of the Head Start group was significantly lower on the Rhyming IGDI than that of the control group (Missall et al. 2006).

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: The developer offers a train-the-trainer model of training support. One member of a team attends a workshop hosted by the developer. The workshop addresses the administration and scoring of the IGDI and the services available from the web site (Missall et al. 2008). The web site provides assessors with information on how to ensure standardized administration of the IGDI. The developer recommends that assessors practice administering the IGDI before official use. In one study, a trained IGDI administrator observed assessors in advance of data collection until the assessors met at least 95 percent of standardized IGDI administration procedures (Missall et al. 2006).

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: The IGDI involves a set of picture cards that must be shuffled and a subset of cards selected for each administration. No two administrations are the same. The administration interval is as frequently as desired, but probably no more than monthly.

Previous Version: None.

NCEE or REL Study Use:[3] Even Start Classroom Literacy Interventions and Outcomes (CLIO)

[1] The U.S. Department of Education, Office of Special Education Programs funded development of the IGDI for children age birth to 8 years.

[2] IGDIs assessing the development of infants and toddlers and early elementary students are also available. Juniper Gardens Children's Project at the University of Kansas (http://www.igdi.ku.edu/index.htm) publishes infant and toddler IGDIs. Early elementary IGDIs, known as Dynamic Indicators of Basic Early Literacy Skills (DIBELS), are published by the Center on Teaching and Learning at the University of Oregon (https://dibels.uoregon.edu/) (see separate profile in the current compendium). In addition, the Get it Got it Go! web site notes that an early numeracy monitoring tool is available from Robin Hojnoski at Lehigh University (roh206@lehigh.edu), and gross motor–related IGDIs are available from the Kinesiology Department at the University of Minnesota (http://cehd.umn.edu/ceed/projects/movement/default.html).

[3] See Table F.1 for web address.

References:

Center for Early Education and Development, University of Minnesota. "Preschool IGDI Frequently Asked Questions." Available at [http://ggg.umn.edu/]. 1998.

Center for Early Education and Development, University of Minnesota. "Preschool IGDI Interpreting Get it Got it Go! Reports." Available at [http://ggg.umn.edu/go/go_interpretreports.html]. 1998.

Center for Early Education and Development, University of Minnesota. "Preschool IGDI Standardized Administration Does Matter." Available at [http://ggg.umn.edu/get/standardization.html]. 1998.

Center for Early Education and Development, University of Minnesota. "Preschool IGDI Using Get it Got it Go! Reports within a Decision-Making Framework: Screening, Testing, and Evaluating with IGDIs." Available at [http://ggg.umn.edu/go/go_decisionmaking framework.html]. 1998.

McConnell, Scott R., Mary McEvoy, Judith J. Carta, Charles R. Greenwood, Ruth Kaminski, Roland H. Good III, Mark Shinn, James Ysseldyke, and Paula Goldberg. "Technical Report #4: Research and Development of Individual Growth Indicators and Development Indicators

for Children between Birth to Age Eight." Minneapolis, MN: Early Childhood Research Institute on Measuring Growth and Development, University of Minnesota, 1998.

McConnell, Scott R., Mary McEvoy, Judith J. Carta, Charles R. Greenwood, Ruth Kaminski, Roland H. Good III, Mark Shinn, James Ysseldyke, and Paula Goldberg. "Technical Report #3: National Survey to Validate General Growth Outcomes for Children Birth to Age Eight—Initial Results." Minneapolis, MN: Early Childhood Research Institute on Measuring Growth and Development, University of Minnesota, 1998.

McConnell, Scott R., Mary McEvoy, Judith J. Carta, Charles R. Greenwood, Ruth Kaminski, Roland H. Good III, Mark Shinn, James Ysseldyke, and Paula Goldberg. "Technical Report #2: Selection of General Growth Outcomes for Children between Birth and Age Eight." Minneapolis, MN: Early Childhood Research Institute on Measuring Growth and Development, University of Minnesota, 1998.

McConnell, Scott R., Jeffrey S. Priest, Shanna D. Davis, and Mary A. McEvoy. "Best Practices in Measuring Growth and Development for Preschool Children." In *Best Practices in School Psychology IV*, edited by Alex Thomas and Jeff Grimes. Bethesda, MD: National Association of School Psychologists, 2002.

Missall, Kristen N. "Reconceptualizing School Adjustment: A Search for Intervening Variables." Unpublished dissertation. Minneapolis, MN: University of Minnesota, 2002.

Missall, Kristen N., Judith J. Carta, Scott R. McConnell, Dale Walker, and Charles R. Greenwood. "Using Individual Growth and Development Indicators to Measure Early Language and Literacy." *Infants and Young Children*, vol. 21, no. 3, 2008, pp. 241-253.

Missall, Kristen N., and Scott R. McConnell. "Technical Report--Psychometric Characteristics of Individual Growth & Development Indicators: Picture Naming, Rhyming, and Alliteration." Minneapolis, MN: Center for Early Education and Development, University of Minnesota, 2004.

Missall, Kristen N., Scott R. McConnell, and Karen Cadigan. "Early Literacy Development: Skill Growth and Relations between Classroom Variables for Preschool Children." *Journal of Early Intervention*, vol. 29, no. 1, 2006, pp. 1-21.

PRESCHOOL LANGUAGE SCALE FOURTH EDITION (PLS-4), 2002

Authors: Irla Lee Zimmerman, Violette G. Steiner, and Roberta Evatt Pond		**Type of Assessment:** Individual assessment **Domain:** Reading (phonological awareness); language arts/language proficiency (assessment of expressive as well as receptive language skills both oral and written, comprehension of basic vocabulary, and grammatical markers)
Publisher: Harcourt Assessment, Inc. 800-211-8378 http://www.harcourtassessment.com		**Grade/Age Range:** Birth to 6 years, 11 months **Administration Interval:** 3-month interval in first year; 6-month interval for older children
Material, Training, and Scoring Costs: PLS-4 English Value Pack with Manipulatives (includes Examiner's Manual, Picture Manual, 15 Record Forms, and 23 Manipulatives): $290 PLS-4 Screening Test Kit (includes Stimulus Book/Test Manual with stimulus pages, technical information, administration and scoring directions, and 25 Record Forms for each age): $135		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2+ (certification beyond bachelor's degree such as a master's degree) **Personnel for Administration:** Highly trained individual **Training for Administration:** Degree or professional experience required (clinician or specialist with experience in diagnostic assessment, such as educational diagnosticians, psychologists, early childhood specialists, or speech-language pathologists)
Languages: English, Spanish		**Alternate Forms:** No
Representativeness of Norming Sample: The PLS-4 was normed in 2002 by using 2000 U.S. Census figures for children age birth through 6 years. The norming sample included 1,534 children (75 to 110 students for each age group, broken down by 2-month intervals in the first year and 6-month intervals thereafter). The sample includes equal numbers of males and females, various ethnic minorities (39.1 percent), children with disabilities (13.2 percent), and bilingual speakers (3.4 percent) from 357 sites in 48 states.		**Summary** **Initial Material Cost:** 3 ($200 to $500) **Time to Administer:** 20 to 45 minutes **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Constructive/Concurrent Validity:** Available **Norming Sample Characteristics:** 3 (normed within the last 10 years and nationally representative)

NARRATIVE

Description: The Preschool Language Scale (PLS-4) is a diagnostic instrument for evaluating language development and identifying language disorders or delays among children from birth through age 6 years, 11 months. It is an individually administered assessment used to measure receptive and expressive skills that are considered to be language precursors. The PLS-4 includes two clusters or scales—Auditory Comprehension and Expressive Communication. The former measures a child's ability to be attentive and respond to stimuli in the environment and to comprehend basic vocabulary or gestures. The Expressive Communication cluster focuses on social communication, expressive language skills, and vocal development. Both clusters include subtests (called tasks), with 4 subtests for each three-month interval for age birth through 11 months and 12 receptive/expressive subtests for each six-month interval for age 1 through 6 years. In total, the PLS-4 contains 68 items, with 2 to 8 items in each subtest. Average administration times vary by age: 20 to 40 minutes for birth to 11 months; 30 to 40 minutes for 1 to 3 years, 11 months; and 25 to 45 minutes for 4 to 6 years, 11 months. The PLS-4 also includes the use of manipulatives (such as a ball, rattle, cups, and crackers) and easel administration (a Picture Manual); the assessor uses the objects or pictures as prescribed to observe the student's reaction or response. The PLS-4 also contains three optional supplemental measures not incorporated into the assessment scores—the Articulation Screener, a Language Sample Checklist (LSC) to evaluate language skills in conversational speech, and, for those children 3-years-old or younger, the Caregiver Questionnaire (CQ) to elicit information on the child's communication behavior at home and the needs of the family.

The publisher's web site also notes a PLS-4 Screening Test that can be used to screen for a broad spectrum of speech and language skills in young children in about 5 to 10 minutes. Paraprofessionals or teachers' aides may administer the PLS-4 Screening Test.

Other Languages: A Spanish version of the PLS-4 (published in 2002) was normed by using a different standardization sample of 1,188 Spanish-speaking children, of whom 81 percent came from homes where Mexican Spanish was spoken. The Spanish version with its own manual is available separately on the publisher's web site.

Uses of Information: The PLS-4 is used to assess language development and to determine whether a child has a language disorder and, if so, whether the source of the disorder is an auditory, expressive, or overall problem.

Methods of Scoring: The assessor records the source/type of response to each item and marks it as correct or incorrect. To obtain a raw score, the assessor sums the items with a 1 (a correct response) and then subtracts the number of incorrect items. A total score as well as Auditory Comprehension and Expressive Communication scores may be obtained. Raw scores may then be converted to standard scores, percentile ranks, and age equivalents.

Interpretability: The Examiner's Manual provides detailed guidelines on the interpretation of scores as related to determining the severity of the disorder and the need for intervention. Specifically, the standard score and percentile ranks help determine the severity of the disorder and identify areas for in-depth testing before defining therapy goals. Using the task analyses (the

PLS-4 Checklist and Profile), a clinician may evaluate the child's strengths, emerging skills, and deficits. The Checklist groups the PLS-4 subtests by age; the Profile groups the subtests by type of language skill tested. The Examiner's Manual recommends that only specialists or clinicians such as speech pathologists, early childhood specialists, psychologists, educational diagnosticians, or other professionals administer, score, and interpret the assessment. Paraprofessional staff, however, may be trained to administer the assessment (details on required time and qualifications are not described).

Reliability:
(1) Internal consistency reliability: For children from birth to 6 years, 11 months, Cronbach's alphas ranged from 0.66 to 0.92 for Auditory Comprehension scores, from 0.73 to 0.95 for Expressive Communication scores, and from 0.81 to 0.97 for the Total Language Score across subtests and age (age groups were split by three-month intervals under age 1 and by six-month intervals thereafter).
(2) Test-retest reliability: Correlations between scores from two administrations (intervals of 2 to 14 days) ranged from 0.85 to 0.91 for Auditory Comprehension, 0.82 to 0.94 for Expressive Communication, and 0.90 to 0.97 for total scores across age groups (divided by five-month intervals) based on 218 2- to 6-year-olds from the standardization sample.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: Certain Expressive Communication items that require open-ended responses. To measure the reliability of ratings on these items, 15 elementary school teachers were trained to score the full assessment and had three weeks of experience in using the PLS-4 scoring rules. Inter-rater reliability was calculated for the full Expressive Communication scale based on 100 protocols each scored by 2 out of 15 scorers, with a resulting correlation of 0.99.

Validity Evidence:
Developers measured relevance and coverage of the content based on a literature review, a user survey, and content reviews. The PLS-3 tasks were modified by using research data, and speech-language pathologists developed new tasks and items. Developers used a "tryout" or pilot to test the PLS-4 on a national sample of 661 children at 227 sites in 46 states. The tryout phase consisted of 229 subtests (tasks) for each age group as well as tasks from the PLS-3. To determine any aspects of the subtests that may affect children inappropriately, an additional 53 children with language disorders were tested. Developers than revised or deleted specific subtests and their subitems once the results were collected from the bias review, statistical analyses, and examiner feedback.

Construct/Concurrent validity: Correlations with the previous version (PLS-3; 2- to 14-day intervals between administrations) were 0.65 for Auditory Comprehension scores and 0.79 for Expressive Communication scores for 104 2- through 6-year-olds. Correlations across subtests or for total scores were not provided.

Four studies examined differences between children who were "typically developing" and children previously identified with (1) a language disorder, (2) a developmental language delay, (3) autism, or (4) a hearing impairment, respectively, in each study. Each study included 60 to 120 students, with equal numbers with or without a disability, who were three to six years old. The standard score means for children with a disability ranged from 64.4 to 78.7 (across all four

studies) while the means for the comparison group were 100 or higher. Sensitivity and specificity were reported for the first study, comparing children with a language disorder to typically developing children only. Sensitivity estimates ranged from 0.77 to 0.80 and specificity estimates ranged from 0.84 to 0.92 across age groups.

Predictive validity: No information available.

Bias Analysis: New or modified tasks were submitted to a panel of experts for two bias reviews to determine appropriateness for children from varied backgrounds (for example, socioeconomic status, ethnicity, and geographic region).

Training Support: No information available.

Adaptations/Special Instructions for Individuals with Disabilities: The Examiner's Manual provides special instructions on administration for children with autism and other severe developmental delays, children with physical impairments (such as hearing and vision), and children who use sign language. For example, the instructions call for different start points depending on level of impairment or disability, splitting up administration of the assessment into short sessions, removing distracting elements from the room, or using gestures/pointing. While the Examiner's Manual provides no special instructions for children who are English Language Learners (ELL), it does offer some instructions for children from "non-mainstream" cultures.

Alternate Forms: None.

Previous Version: The PLS-4 is similar to the PLS-3 in its design, subscales, and overall skills assessed. The PLS-3 used standardization data from 1980 U.S. Census figures while the PLS-4 has been updated to use standardization data based on 2000 U.S. Census figures and a more diverse sample. In addition, the PLS-4 has been revised to extend the age-appropriateness of the scale for children from birth to 11 months and those 5 through 6 years, 11 months. The Auditory Comprehension and Expressive Communication scales are now grouped into four two-month subtests for children birth to 11 months and four five-month subtests for children age 5 through 6 years, 11 months. According to the developers, the age intervals were revised for the PLS-4 so that the skills of youngest and oldest students could be better assessed and the floor and ceiling of the assessment could be improved.

NCEE or REL Study Use:[1] National Evaluation of Early Reading First

[1] See Table F.1 for web address.

References:

Flowerday, Terri. "Review of Preschool Language Scale, Fourth Edition." In *16th: The Sixteenth Mental Measurements Yearbook*, edited by Robert A. Spies and Barbara S. Plake. Lincoln, NE: The Buros Institute of Mental Measurements, 2005.

Kisker, Ellen E., Kimberly Boller, Charles Nagatoshi, Christine Sciarrino, Vinita Jethwani, Teresa Zavitsky, Melissa Ford, and John M. Love. "Resources for Measuring Services and Outcomes in Head Start Programs Serving Infants and Toddlers." Washington, DC: Mathematica Policy Research, 2003.

Suen, Hoi K. "Review of Preschool Language Scale, Fourth Edition." In *16th: The Sixteenth Mental Measurements Yearbook*, edited by Robert A. Spies and Barbara S. Plake. Lincoln, NE: The Buros Institute of Mental Measurements, 2005.

Zimmerman, Irla L., Violette G. Steiner, and Roberta E. Pond. *Preschool Language Scale 4: Picture Manual*. San Antonio, TX: Harcourt Assessment, Inc., 2002.

Zimmerman, Irla L., Violette G. Steiner, and Roberta E. Pond. *Preschool Language Scale, Fourth Edition, Examiner's Manual*. San Antonio, TX: Harcourt Assessment, Inc., 2002.

Zimmerman, Irla L., Violette G. Steiner, and Roberta E. Pond. *Preschool Language Scale, Fourth Edition (PLS-4) Spanish Edition*. San Antonio, TX: Harcourt Assessment, Inc., 2002.

THE RESEARCH ASSESSMENT PACKAGE FOR SCHOOLS-STUDENT SELF REPORT (RAPS-S), 1998

Authors: Institute for Research and Reform in Education		**Type of Assessment:** Student self-report (group administered) **Domains:** Approaches to learning/motivation
Publisher: Institute for Research and Reform in Education 732-557-0200 http://www.irre.org		**Grade/Age Range:** Grade 3 through 8 **Administration Interval:** None described
Material, Training, and Scoring Costs: The manual and questionnaire are available at no cost on the developer's web site. The developer must be cited if materials are copied for distribution.		**Personnel and Training Requirements** **Credentials Required for Use:** Level 1 (training or supervised experience with measurement) **Personnel for Administration:** Highly trained individual **Training for Administration:** Minimal (1 to 2 hours) Assessors should be trained to respond uniformly to student questions.
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (under $100) **Time to Administer:** 50 minutes **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (meets minimum acceptability ratings—0.70) **Predictive Validity:** Available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Research Assessment Package for Schools (RAPS) is a multitool assessment that includes five measures with varying reporters (students, teachers, parents, and school records).[1] The RAPS student self-report (RAPS-S) is a questionnaire administered to students that measures three subscales (termed "domains" by developers): level of engagement in school (i.e., Engagement), beliefs about themselves (i.e., Beliefs), and the interpersonal supports they receive from parents and teachers (i.e., Support). To encourage honest responses from students, teachers are advised not administer the questionnaire to their own students. Seven domains are embedded in the three subscales. Separate RAPS-S questionnaires are used for elementary school students (grade 3 through 5) and middle school students (grade 6 through 8). Students complete the paper-and-pencil questionnaire by circling responses for 88 items (elementary school form) or 84 items (middle school form). The four Likert-type scale response options available on the RAPS-S are "very true," "sort of true," not very true," and "not at all true." One item, assessing the importance of doing one's best at school, contains a similar scale of four responses, ranging from "very important" to "not at all important." The RAPS-S takes approximately 50 minutes to administer.

Other Languages: None.

Uses of Information: The RAPS-S may be used as a diagnostic assessment for all but high-risk students to provide data on levels of student engagement, student self-beliefs, and perceptions of support in school. Data may be aggregated at the school or district level. The assessment may also be used to evaluate changes in reports of engagement, beliefs, and perceptions of support. Furthermore, optimal or high-risk RAPS-S scores may be used to predict successful or poor student academic performance.

Methods of Scoring: RAPS-S scores for each subscale (Engagement, Beliefs, and Support) are made up of other composite scores, which consist of unique subdomains and constructs. The RAPS-S has 17 composite scores for the elementary school student questionnaire and 19 composite scores for the middle school student questionnaire. The manual provides instructions for which items to include in each subscale as well as the formulas needed to calculate composite and total scores. The manual does not specify who should conduct the scoring.

Interpretability: Continuous RAPS-S scores are classified as indicators of high risk, other, and optimal student performance based on various cut points provided in the manual. For example, RAPS-S Engagement scores less than 3.25 are deemed high-risk; scores equal to or greater than 3.25 and less than 3.75 are classified as other; and scores equal to or greater than 3.75 are optimal. Students with scores in the optimal and high-risk range are assigned an indicator score of 1, and students with scores in the middle other range are assigned an indicator score of 0.

Reliability:
(1) Internal consistency reliability: Developers reported Cronbach's alpha coefficients for raw scores from the RAPS-S subscales of 0.71 (Engagement), 0.87 (Beliefs), and 0.87 (Support) among elementary school students and of 0.77, 0.87, and 0.88, respectively, among middle school students. Estimates were based on a convenience sample of 1,800 students from six elementary schools in one urban district and 2,400 students in three middle schools in an urban

and a suburban district (date not specified). The majority of students were Black and eligible for free or reduced-price school lunch (in urban locales). Several years of student data were available from four of the elementary schools and all of the middle schools.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: Not applicable.

Validity Evidence:
The RAPS assessments belong to a larger school site reform framework designed to promote the achievement of success. Developers cited previous research and theoretical work related to the framework (IRRE 2003; Connell and Klem 2002). The RAPS-S subscales are based on a literature review of conditions that are expected to enhance student academic performance.

Construct/Concurrent validity (see sample information under Internal consistency reliability): Developers correlated scores on the RAPS-S, the RAPS-T (a brief teacher questionnaire regarding each of their students' level of engagement), and the Student Performance and Commitment Index (SPCI)[2] by RAPS-S subscale and school level. They used point-biserial coefficient estimates as correlations between RAPS-S continuous subscale scores with the dichotomous indicator variables of (1) optimal versus other, (2) high risk versus other,[3] and (3) optimal versus high risk on the SPCI, RAPS-S, and RAPS-T. Although not shown here, the manual provides phi coefficient estimates such that only dichotomous RAPS-S, RAPS-T, and SPCI indicator scores were compared.

Developers provide intercorrelations between RAPS-S subscales. Coefficients between continuous Beliefs scores and dichotomous Engagement scores correlated from 0.42 to 0.63 among elementary school students and from 0.27 to 0.67 among middle school students. Coefficients for continuous Support and dichotomous Engagement scores ranged from 0.45 to 0.67 and from 0.39 to 0.69 for elementary and middle school students, respectively, and correlations between continuous Support and dichotomous Beliefs scores ranged from 0.40 to 0.74 and from 0.37 to 0.69.

With respect to correlations between continuous RAPS-S subscale scores and dichotomous RAPS-T scores, coefficients between RAPS-S Engagement scores and RAPS-T scores ranged from 0.22 to 0.36 and from 0.22 to 0.42 for elementary and middle school students, respectively. RAPS-S Beliefs scores correlated with the RAPS-T from 0.17 to 0.27 (elementary school) and from 0.16 to 0.32 (middle school). RAPS-S Support scores correlated with the RAPS-T from 0.20 to 0.31 (elementary school) and from 0.15 to 0.31 (middle school).

In regard to correlations between continuous RAPS-S subscale scores and dichotomous SPCI scores, coefficients between RAPS-S Engagement and SPCI scores ranged from 0.13 to 0.20 and from 0.14 to 0.25 for elementary and middle school students, respectively. RAPS-S Beliefs scores correlated with SPCI scores from 0.14 to 0.23 (elementary school) and from 0.12 to 0.22 (middle school). RAPS-S Support scores and SPCI scores correlated from 0.14 to 0.24 (elementary school) and from 0.09 to 0.18 (middle school).

Predictive validity: Continuous RAPS-S Engagement scores were correlated with dichotomous SPCI scores assessed later in time for student subsamples that had both RAPS and SPCI data.

Elementary school students' RAPS-S Engagement scores correlated with their SPCI scores in middle school from 0.14 to 0.24. Middle school students' RAPS-S Engagement scores correlated with their SPCI scores in high school from 0.06 to 0.10.

Bias Analysis: No information available.

Training Support: The manual provides instructions for oral administration of the RAPS-S regarding who administers group administrations, standardization of instructions, question pacing, and when to read response options. Administrators must also be trained to respond to questions from students.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: None.

Previous Version: None.

NCEE or REL Study Use:[4] Use of Classroom Assessment for Student Learning (REL-Central)

[1] The RAPS contains five tools. In addition to the RAPS-S profiled here, the RAPS-T is a brief questionnaire (three items) administered to teachers to determine each student's level of engagement; the RAPS-P is a questionnaire administered to parents regarding their children and teachers; the RAPS-R is a system for analyzing data from student records; and the T-RAPS is a questionnaire administered to teachers to determine their level of engagement and perceived professional and interpersonal supports at school. The manual does not present information on the RAPS-R, RAPS-P, and the T-RAPS. At the time of publication (1998), the RAPS-P and T-RAPS were undergoing further field testing and psychometric work.

[2] SPCI dichotomous scores were based on attendance records and standardized achievement assessments in reading and mathematics.

[3] All point-biserial correlation coefficients comparing continuous RAPS scores to the comparison measure's high risk versus other indicator scores are negative, demonstrating that lower RAPS scores are associated with student high risk. For ease of reporting, ranges of point-biserial coefficients are shown in absolute value regardless of direction of the comparison. Appendices in the manual provide individual negative values.

[4] See Table F.1 for web address.

References:

Connell, James P., and Adena M. Klem. "A Theory-of-Change Approach to Evaluating Investments in Public Education." In *Measuring the Impact of the Nonprofit Sector*, edited by Patrice Flynn and Virginia A. Hodgkinson. New York: Kluwer Academic/Plenum Publishers, 2002.

Institute for Research and Reform in Education (IRRE). "First Things First: A Framework for Successful School Reform." Submitted to the Ewing Marion Kauffman Foundation. Philadelphia: Institute for Research and Reform in Education, 2003.

Institute for Research and Reform in Education (IRRE). *Research Assessment Package for Schools (RAPS) Manual*. Philadelphia: Institute for Research and Reform in Education, 1998.

SCIENCE READING COMPREHENSION ASSESSMENT, 2007

Authors: Educational Testing Service (ETS)		**Type of Assessment:** Group-administered assessment **Domain:** Reading (comprehension)
Publisher: Educational Testing Service 609-921-9000 http://www.ets.org		**Grade/Age Range:** Grade 5 **Administration Interval:** Annual[1]
Material, Training, and Scoring Costs: Not specified		**Personnel and Training Requirements** **Credentials Required for Use:** Not specified **Personnel for Administration:** Individual with basic clerical skills with some training[1] **Training for Administration:** Minimal (1 to 2 hours)[1]
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: The norming sample consisted of 2,921 students (47 percent male and 46 percent female, and 7 percent unaccounted for) in grade 5. The assessment was administered in 2007. The sample was not nationally representative, but it spanned 89 schools in a geographically diverse area.		**Summary** **Initial Material Cost:** To be determined upon negotiation with the publisher **Time to Administer:** About 45 to 50 minutes[1] **Ease of Administration and Scoring:** 4 (administered or scored by a clinician or specialist) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 2 (older than 10 years or not nationally representative)

NARRATIVE

Description: The Science Reading Comprehension Assessment is a group-administered assessment to determine grade 5 students' comprehension of expository text in science. Students read passages and complete multiple-choice items aimed at assessing general reading comprehension skills (for example, vocabulary, main idea identification, inference, purpose) from science-based passages. Students are not expected to bring a certain level of science knowledge to the assessment; answers regarding specific scientific knowledge must be gleaned from the passage. The paper-and-pencil assessment comprises five passages with 30 multiple-choice items in total. It is untimed but is designed to take from 45 to 50 minutes for completion (N. Carey, personal communication, December 16, 2008). The Science Reading Comprehension Assessment was developed in particular for the NCEE's Evaluation of Reading Comprehension Programs, which focused on grade 5 students.

Other Languages: None.

Uses of Information: The Science Reading Comprehension Assessment may be used to assess grade 5 students' ability to comprehend expository text focused on science content. In particular, the Evaluation of Reading Comprehension Programs used the measure for evaluating such reading interventions.

Methods of Scoring: Each item answered correctly receives a score of one point. Total raw scores are the sum of correct items. The developers used Item Response Theory (IRT) to estimate theta scores based on the three-parameter logistic IRT model. The IRT theta scores were then linearly transformed to a reporting scale with a mean of approximately 500 and a standard deviation of 30. ETS conducts the scoring. The Technical Report provides tables to convert the raw score into a scale score.

Interpretability: A higher score means a higher level of ability to comprehend science-based reading passages. The IRT scale scores are based on the current sample of grade 5 students.

Reliability:
(1) Internal consistency reliability: Cronbach's alpha for raw scores was 0.85 (N = 2,912), calculated from the norming sample but excluding those who responded to fewer than five items.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: No information available.

Validity Evidence:
To determine their inclusion in the assessment, passages and items underwent review by three content experts, an editor (for clarity, spelling, grammar, style), and a researcher (for factual accuracy). In addition, the developers selected passages according to the following criteria: a science focus (inclusion of at least one passage focused on a scientific principle or concept); readability at a grade 4 and 5 level; and reflective of interests of grade 5 students. The following criteria guided the selection of questions about the passages: assessment of general standards of vocabulary, main idea, details, inference and purpose; vocabulary and sentence structure appropriate for grade level; and items forming a distribution in difficulty with 25 percent of low

difficulty, 50 percent medium, and 25 percent high. Experts from both ETS and Mathematica reviewed the passages and items, and the assessment was then piloted.

Construct/Concurrent validity: The developers conducted Classical Item Analysis and flagged items for review on the basis of five characteristics: item difficulty *p*-value less than 0.25 or greater than 0.90; a correlation between a correct item response and total test performance less than 0.15; a correlation between an incorrect item response and total test performance greater than 0; an item omission rate of 5 percent or greater; or a lack of selection of certain options on an item. Three items were flagged for review by a content expert and a psychometrician. After the three items were reviewed and bias analysis was conducted (see Bias Analysis), it was determined that all items were valid. The developers used all items in IRT calibrations for scoring.

Correlation between students' scores on the Science Reading Comprehension Assessment and the Group Reading Assessment and Diagnostic Evaluation (GRADE) in the spring was 0.70. In addition, correlation between class-level means on the Science Reading Comprehension Assessment and the Social Science Reading Comprehension Assessment (a similar assessment using expository text on social science topics) was 0.77 for approximately 270 classrooms. (Student-level scores on both the Science and Social Science assessments were unavailable for a given student because students were randomly assigned to take only one of the comprehension assessments. However, student-level scores were available to permit correlation with the GRADE. These class-level correlations might be higher than what would be expected at the student-level given that the class-level correlation with the GRADE was higher than the student-level correlation noted above.)

Predictive validity: No information available.

Bias Analysis: During test development, an ETS staff member trained in issues of fairness reviewed potential items for inclusion, devoting special attention to issues of language use, offensive content, and racial/ethnic or gender bias. After the assessment was piloted, the developers conducted differential item functioning (DIF) analyses based on gender and race/ethnicity. Race/ethnicity analyses focused on White, Black, and Hispanic students because sample sizes for other ethnic groups were too small. The analysis found no items exhibiting DIF across subgroups.

Training Support: A two-hour training session reviews the goal of the test and basic proctoring information as well as how to read instructions verbatim and how to respond to students' questions (N. Carey, personal communication, December 16, 2008).

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: None.

Previous Version: None.

NCEE or REL Study Use:[2] Evaluation of Reading Comprehension Programs

[1] N. Carey, personal communication, December 16, 2008.

[2] See Table F.1 for web address.

References:

Educational Testing Service. "Mathematica Reading Comprehension Assessments 2007 Technical Report." Princeton, NJ: Educational Testing Service, 2007a.

Educational Testing Service. *Science Reading Comprehension Assessment* (unpublished). Princeton, NJ: ETS, 2007b.

James-Burdumy, Susanne, David Myers, John Deke, Wendy Mansfield, Russell Gersten, Joseph Dimino, Jan Dole, Lauren Liang, Sharon Vaughn, and Meaghan Edmonds. "The National Evaluation of Reading Comprehension Interventions: Design Report, Final Report." Submitted to the Institute of Education Sciences, U.S. Department of Education,. Princeton, NJ: Mathematica Policy Research, 2006.

James-Burdumy, Susanne, Wendy Mansfield, John Deke, Nancy Carey, Julieta Lugo-Gil, Alan Hershey, Aaron Douglas, Russell Gersten, Rebecca Newman-Gonchar, Joseph Dimino, and Bonnie Faddis. "Effectiveness of Selected Supplemental Reading Comprehension Interventions: Impacts on a First Cohort of Fifth-Grade Students." (NCEE 2009-4032). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

SELF- AND TASK-PERCEPTION QUESTIONNAIRE, 1995

Authors: Jacquelynne S. Eccles and Allan Wigfield		**Type of Assessment:** Student self-report **Domain:** Approaches to learning/ motivation (mathematics-specific)
Publisher: Unpublished; items and response format list Eccles and Wigfield (1995)		**Grade/Age Range:** Grades 5 through 12 **Administration Interval:** None described
Material, Training, and Scoring Costs: Not specified		**Personnel and Training Requirements** **Credentials Required for Use:** Not specified **Personnel for Administration:** Individual with basic clerical skills and some training **Training for Administration:** Self-training (<1 hour)
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (<$100) **Time to Administer:** Not specified **Ease of Administration and Scoring:** 1 (not described) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Self- and Task-Perception Questionnaire is a student self-report of adolescent values, attitudes, and beliefs about mathematics achievement. It has been used with students in grades 5 through 12 and may be group-administered. The questionnaire includes 19 items divided into three subscales: Perceived Task Value (7 items), Ability/Expectancy (5 items), and Perceived Task Difficulty (7 items). The Perceived Task Value subscale is further divided into three components: Intrinsic Interest Value (or enjoyment of mathematics activities, with 2 items), Perceived Attainment Value/Importance (3 items), and Extrinsic Utility Value (or usefulness of mathematics for achieving future goals, with 2 items). The Perceived Task Difficulty subscale is divided further into two components: Task Difficulty (3 items) and Required Effort (4 items). Students rate each item (e.g., "I find working on math…") for level of agreement, choosing from one of seven responses on a Likert-type scale that ranges from "very boring" to "very interesting."

Other Languages: None.

Uses of Information: The Self- and Task-Perception Questionnaire may be used to measure constructs related to motivation for student achievement. The questionnaire contrasts previous theorists' emphasis on Ability/Expectancy-related items by increasing the focus on task-related items. Developers assert that three separate dimensions of motivation—Perceived Task Value, Ability/Expectancy, and Perceived Task Difficulty—are distinguishable from one another and may be considered unique constructs. Several studies have used variations of the Self- and Task-Perception Questionnaire to assess self-perceptions about mathematics, reading, music, and physical activity (Wigfield et al. 1997; Kellow and Jones 2005; Sabiston and Crocker 2008).

Methods of Scoring: Responses for each of the 19 items are expressed on a 7-point Likert-type scale. Score calculation was not described.[1]

Interpretability: Guidelines for interpreting scores are not readily available.

Reliability:
(1) Internal consistency reliability: Developers presented Cronbach's alpha coefficients for scores from each subscale and for scores of components within subscales. Within the Perceived Task Value subscale, reliability estimates for scores on the Intrinsic Interest Value, Attainment Value/Importance, and Extrinsic Utility Value components were 0.76, 0.70, and 0.62, respectively. Scores for the Ability/Expectancy subscale had a reliability estimate of 0.92. Within the Perceived Task Difficulty subscale, reliability estimates for scores on the Task Difficulty and Required Effort components were 0.80 and 0.78, respectively. The student sample comprised 707 White adolescents from middle-class homes in grades 5 through 12, of whom roughly half were female (year and location not specified).

A separate investigation for the Ability/Expectancy subscale reported a Cronbach's alpha coefficient of 0.85 (Kellow and Jones 2005). The student sample included 81 Black and White grade 9 students from a Florida school. Another study (Wigfield et al. 1997) that used a modified version of the Self- and Task-Perception Questionnaire provided reliability estimates for scores on Competence Belief, Usefulness-Importance, and Perceived Interest (adaptations of the

Perceived Task Value and Ability/Expectancy subscales). Subscale adaptations included item re-ordering, re-wording, and removal (two items) and the addition of three new items. Reliability estimates for scores of these subscales ranged from 0.61 to 0.92 to include all measured domains (mathematics, reading, music, and sports), with reliability estimates for mathematics subscales not specified. The student sample included 615 students in grades 1 through 6 from four school districts of a large Midwestern city.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: Not applicable.

Validity Evidence:
Developers of the Self- and Task-Perception Questionnaire conducted exploratory factor analyses to test the three-subscale structure and to assess the number of factors within each subscale (Eccles and Wigfield 1995). The sample comprised 707 White, middle-class adolescents in grades 5 through 12, of whom roughly half were female. They analyzed 29 original items (9 items in Perceived Task Value, 10 items in Ability/Expectancy, and 10 items in Perceived Task Difficulty). Results demonstrated three factors in the Perceived Task Value subscale and two factors in the Perceived Task Difficulty subscale. They removed 10 items that did not load highly for any factor within its respective subscale (loading values not reported).

Construct/Concurrent validity: Developers conducted a principal components analysis across the final 19 items, and the results supported the grouping of items into three subscales. Confirmatory factor analysis on the 19 items supported the factor structure developed during exploratory factor analysis—three separate subscales comprised of factors (Eccles and Wigfield 1995).

Kellow and Jones (2005) correlated scores from the Ability/Expectancy subscale with total test scores from the Florida Comprehensive Assessment for mathematics. The correlation coefficient was 0.40.

Predictive validity: No information available.

Bias Analysis: Developers assessed item invariance by gender and age (grades 5 through 7 versus 8 through 12) by using structural equation modeling with maximum likelihood estimation. Results showed reasonable invariance across groups based on goodness-of-fit indexes of 0.96 and 0.95 for gender and age, respectively.

Related studies on elementary school students have shown gender and age differences in mathematics ratings when used with adapted subscales of the Self- and Task-Perception Questionnaire. Wigfield et al. (1997) assessed students' Competence Beliefs, Usefulness-Importance, and Perceived Interest (variations of Perceived Task Value and Ability/Expectancy). Boys had higher Competence Beliefs than girls based on repeated measures MANOVAs. No gender differences existed for the Usefulness-Importance and Perceived Interest subscales. Eccles et al. (1993) assessed differences by grade and gender for mathematics Competence Beliefs and Subjective Task Value (variations of Perceived Task Value, Ability/Expectancy, and Perceived Task Difficulty). Students in grade 1 had higher Competence Beliefs than students in grade 4, and boys had significantly higher Competence Beliefs than girls. Subjective Task Value ratings did not differ by grade or gender.

Training Support: No information available.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: None.

Previous Version: The Self- and Task-Perception Questionnaire is based on items and scales previously developed by Eccles et al. (1984) to test whether gender differences existed for 200 students in grades 8 through 10 in mathematics value perceptions (similar to Perceived Task Value), self concepts of ability (similar to Ability/Expectancy), and perception of the difficulty of mathematics (similar to Perceived Task Difficulty).

NCEE or REL Study Use:[2] The Effect of Connected Mathematics Program 2 (CMP2) (On the Math Achievement of Middle School Students in Selected Schools in the Mid-Atlantic Region) (REL-Mid-Atlantic)

[1] Self- and Task-Perception Questionnaire developers evaluated whether adolescents' perceptions of task value, ability, and task difficulty and the components within each dimension are distinguishable from each other. Developers focused on reporting findings from exploratory and confirmatory analysis to define dimensions more precisely and assess relationships between them.

[2] See Table F.1 for web address.

References:

Eccles (Parsons), Jacquelynne, Terry Adler, and Judith L. Meece. "Sex Differences in Achievement: A Test of Alternate Theories." *Journal of Personality and Social Psychology*, vol. 46, no. 1, 1984, pp. 26-43.

Eccles, Jacquelynne S., and Allan Wigfield. "In the Mind of the Actor: The Structure of Adolescents' Achievement Task Values and Expectancy-Related Beliefs." *Personality and Social Psychology Bulletin*, vol. 21, no. 3, 1995, pp. 215-225.

Eccles, Jacquelynne S., Allan Wigfield, Rena D. Harold, and Phyllis C. Blumenfeld. "Age and Gender Differences in Children's Achievement Self-Perceptions during the Elementary School Years." *Child Development*, vol. 64, 1993, pp. 830-847.

Kellow, J. Thomas, and Brett D. Jones. "Stereotype Threat in African-American High School Students: An Initial Investigation." *Current Issues in Education*, vol. 8, no. 20, 2005, Available at [http://cie.asu.edu/volume8/number20/index.html].

Sabiston, Catherine M., and Peter R.E. Crocker. "Exploring Self-Perceptions and Social Influences as Correlates of Adolescent Leisure-Time Physical Activity." *Journal of Sport and Exercise Psychology*, vol. 30, 2008, pp. 3-22.

Wigfield, Allan, Jacquelynne S. Eccles, Kwang S. Yoon, Rena D. Harold, Amy J.A. Arbreton, Carol Freedman-Doan, and Phyllis C. Blumenfeld. “Change in Children's Competence Beliefs and Subjective Task Values across the Elementary School Years: A 3-Year Study.” *Journal of Educational Psychology*, vol. 89, no. 3, 1997, pp. 451-469.

SOCIAL COMPETENCE AND BEHAVIOR EVALUATION, PRESCHOOL EDITION (SCBE), 1995

Authors: Peter J. LaFreniere and Jean E. Dumas		**Type of Assessment:** Teacher report **Domain:** Social-emotional (behavior, social competence, affective expression, adjustment)
Publisher: Western Psychological Services 800-648-8857 http://www.wpspublish.com		**Grade/Age Range:** 30 to 78 months **Administration Interval:** None described
Material, Training, and Scoring Costs: Kit (includes manual and 25 Autoscore forms): $92.50		**Personnel and Training Requirements** **Credentials Required for Use:** No special qualifications required **Personnel for Administration:** Individual with basic clerical skills with some training **Training for Administration:** Self-training (<1 hour)
Languages: English, French, Spanish		**Alternate Forms:** No
Representativeness of Norming Sample: Normed in the early 1990s (year not specified), the sample included 1,263 students age 30 to 78 months from six sites in Indiana and Colorado. As compared to 1991 U.S. Census data, the norming sample contains a higher proportion of low socioeconomic status parents than the nation. Sample student ages include about 8 percent 3-year-olds, 28 percent 4-year-olds, 42 percent 5-year-olds, and 22 percent 6-year olds. Equal numbers of males and females were included, and the sample was ethnically diverse, including Black, Hispanic, and Asian students.		**Summary** **Initial Material Cost:** 1 (under $100) **Time to Administer:** 15 minutes **Ease of Administration and Scoring:** 2 (self-administered or administered by someone with basic clerical skills) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 2 (older than 10 years or not nationally representative)

NARRATIVE

Description: The Social Competence and Behavior Evaluation (SCBE) is a teacher report of student behavior related to social competence, affective expression, and adjustment; it is used to assess students' social interactions with peers and adults. The assessment is appropriate for students age 30 through 78 months. It is an 80-item questionnaire designed to be completed by the teacher in about 15 minutes. Behaviors are rated on a 6-point scale for frequency of occurrence, ranging from "almost never" to "almost always occurs". The SCBE consists of the eight following subscales, with 10 items each (5 pertaining to successful adjustment and 5 describing adjustment difficulties): Depressive-Joyful, Anxious-Secure, Angry-Tolerant, Isolated-Integrated, Aggressive-Calm, Egotistical-Prosocial, Oppositional-Cooperative, and Dependent-Autonomous. The eight subscales form four summary scales: Social Competence (20 items), Internalizing Problems (20 items), Externalizing Problems (20 items), and General Adaptation (all 80 items).

The SCBE-30 is the short version of the SCBE. It consists of three 10-item subscales that correspond to three summary scales on the long form: Social Competence, Anger-Aggressive (Externalizing Problems), and Anxiety-Withdrawal (Internalizing Problems). Validation evidence for scores of the short forms was obtained by correlating scores from the three 20-item SCBE summary scales (Social Competence, Internalizing Problems, and Externalizing Problems) with scores from the respective counterpart 10-item scales on the SCBE-30. Pearson correlations ranged from 0.92 to 0.97 across subscales.

Other Languages: The SCBE was previously entitled the Preschool Socio-Affective Profile (PSP) (developed as a measure for French speakers), which was pilot tested on a sample of 608 French-Canadian students from 60 preschools in Montreal (LaFreniere et al. 1992). The manual discusses the psychometric properties of an earlier version of the SCBE (also in French), which was pilot-tested on a sample of 979 preschool students in Montreal. Respondents were almost exclusively preschool teachers. Reliability estimates for scores, using the Cronbach's alpha formula, ranged from 0.79 to 0.91 across all eight subscales. In addition, developers estimated levels of inter-rater reliability, which ranged from 0.72 to 0.89 across the eight subscales. Developers also estimated levels of reliability for scores by using the test-retest method, based on a subsample of 29 students over the course of a two-week interval between the first and second test administrations. Correlations ranged from 0.74 to 0.87 across the eight subscales. Factor analysis uncovered the same three factors (Social Competence, Externalizing Problems, and Internalizing Problems) as the subsequent English form of the SCBE.

A Spanish-translated version of the SCBE has been developed and pilot-tested in the United States and Spain (Dumas et al. 1998). In an initial study comparing the English and Spanish versions, conducted in Miami with 159 preschool students ranging in age from 3 to 7 years, internal consistency reliability estimates for scores ranged from 0.80 to 0.88 for the eight subscales on the Spanish version and from 0.77 to 0.83 for the eight subscales on the English version. Internal consistency reliability estimates for scores on the four summary scales ranged from 0.81 to 0.90 on the Spanish version and from 0.77 to 0.93 on the English version. Mean scores for the sample taking the Spanish version were of similar ranges across genders. The Spanish form was standardized by using a sample of 414 preschoolers in Houston and Valencia, Spain, with student ages ranging from 2 years, 5 months to 6 years. Internal consistency

reliability estimates for scores ranged from 0.67 to 0.89 across all sub- and summary scales (breakdown by scale not provided). Factor analysis on the Spanish version produced the same three factors: Social Competence, Externalizing Problems, and Internalizing Problems as the English measure. Using a subsample of 66 students, developers also evaluated levels of test-retest reliability for scores from the Spanish version with a two-week interval between administrations. Pearson correlations ranged from 0.77 to 0.90 for scores of the eight subscales and from 0.73 to 0.88 for scores of the four summary scales.

Uses of Information: The purpose of the SCBE is to assess emotional status, behaviors, and growth in the classroom setting.

Methods of Scoring: The SCBE requires 10 minutes to score. The teacher circles a value from 1 to 6, with anchors at 4 points: 1 = almost never occurs, 3 = sometimes occurs, 5 = often occurs, and 6 = almost always occurs. The assessor transfers the ratings onto the Scoring Sheet, summing the items per the scoring instructions provided in the manual in order to calculate the raw scores for the eight subscales and three summary scales and for a total score—General Adaptation. The scores are then converted to normalized *T*-scores and percentiles by plotting them on the provided Profile sheet. Specific qualifications for scorers are not provided.

Interpretability: The General Adaptation summary score indicates a student's level of adjustment, with higher scores indicating better overall adjustment to the classroom. The manual provides case studies to aid in interpretation.

Reliability:
(1) Internal consistency reliability: Cronbach's alpha coefficients based on the Colorado (N = 439) and Indiana (N = 824) samples ranged from 0.80 to 0.89 for scores across the eight scales.
(2) Test-retest reliability: None described for the English SCBE (see Other Languages for information on the SCBE French and Spanish versions).
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: Based on calculation of different teachers' agreement on independent evaluations of the same child at the same time, inter-rater reliability estimates were obtained for the Indiana sample only (N = 824), and ranged from 0.73 to 0.89 across the eight scales (see Other Languages for information on the SCBE French version, from which the SCBE English was translated).

Validity Evidence:

Construct/Concurrent validity: Principal components analysis (PCA) showed that both Indiana and Colorado samples yielded the same three factors—Social Competence, Externalizing Problems, and Internalizing Problems.

The SCBE was compared with the Child Behavior Checklist (CBCL) for a sample of 177 French-Canadian students. Pearson's correlation coefficients were calculated separately for boys and girls across five subscales for both assessments. Correlations between scores for similar scales across the two assessments were (1) SCBE Anxious with CBCL Anxiety: 0.40 for girls and 0.48 for boys; (2) SCBE Isolated with CBCL Withdrawal: 0.53 for girls and 0.58 for boys; (3) SCBE Aggressive with CBCL Aggression: 0.63 for girls and 0.53 for boys; (4) SCBE

Internalizing Problems with CBCL Internalizing: 0.53 for girls and 0.63 for boys; and (5) SCBE Externalizing Problems with CBCL Externalizing: 0.66 for girls and 0.64 for boys.

Correlations between measures assessing different traits yielded the following coefficients: The SCBE Anxious subscale scores were significantly correlated with the CBCL Withdrawal, Internalizing, and Externalizing subscale scores for girls, with correlations ranging from 0.19 to 0.43; and with the CBCL Withdrawal and Internalizing subscale scores for boys at 0.48 and 0.52, respectively. The subscale was not significantly correlated with the CBCL Aggression subscale scores for girls or boys or with the CBCL Externalizing subscale scores for boys. The SCBE Isolated subscale scores were correlated with the CBCL Anxiety and Internalizing subscale scores at 0.30 and 0.47 for girls, respectively, and at 0.51 and 0.59 for boys, respectively. The scale was not significantly correlated with the CBCL Aggression or Externalizing subscales. The SCBE Aggressive subscale scores were significantly correlated with the CBCL Externalizing subscale scores at 0.61 for girls and 0.49 for boys but were not significantly correlated with the CBCL Anxiety, Withdrawal, or Internalizing subscale scores. The SCBE Internalizing Problems subscale scores were significantly correlated with the CBCL Anxiety, Withdrawal, Aggression, and Externalizing subscale scores for girls, with correlations ranging from 0.20 to 0.50. The scale also correlated with CBCL Anxiety, Withdrawal, and Externalizing subscales for boys (but not with CBCL Aggression), with correlations ranging from 0.27 to 0.60. The SCBE Externalizing Problems subscale scores were significantly correlated with the CBCL Withdrawal and Aggression subscale scores for girls, at -0.20 and 0.71, respectively, but was not significantly correlated with the CBCL Anxiety or Internalizing subscale scores for girls. It was significantly correlated with the CBCL Aggression subscale for boys at 0.78 but not significantly correlated with the CBCL Anxiety, Withdrawal, or Internalizing subscales for boys.

Two criteria were used for conducting discriminant analyses—peer sociometric ratings and direct observation of social participation (50 focal-child one-minute samples during free play over one month). Teachers completed the SCBE for 126 randomly selected students from a larger sample of 994 students in Montreal. Based on test results, the students were organized into four groups: socially competent, anxious-withdrawn, angry-aggressive, and average. One-way ANOVAs were calculated for the four comparison groups on their sociometric ratings and observed play. The SCBE scales differentiated the anxious-withdrawn and the angry-aggressive groups on these measures. The SCBE anxious-withdrawn group spent significantly more time in non-interaction (34 percent) than the angry-aggressive group (18 percent). The angry-aggressive group received the most peer rejections or negative nominations. The socially competent group received the most positive nominations and the least negative nominations compared to all other groups.

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: No information available.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: None.

Previous Version: The SCBE's predecessor was the Preschool Socio-Affective Profile (PSP), administered in French. No specifics are provided on any changes.

NCEE or REL Study Use:[1] National Evaluation of Early Reading First

[1] See Table F.1 for web address.

References:

Berry, Daniel J., Lisa J. Bridges, and Martha J. Zaslow. "Early Childhood Measures Profiles." Washington, DC: Child Trends, 2004.

Dumas, Jean E., Alfonso Martinez, and Peter J. LaFreniere. "The Spanish Version of the Social Competence and Behavior Evaluation (SCBE)--Preschool Edition: Translation and Field Testing." *Hispanic Journal of Behavioral Sciences*, vol. 20, no. 2, 1998, pp. 255-268.

LaFreniere, Peter J., and Jean E. Dumas. "Social Competence and Behavior Evaluation in Children Ages 3 to 6 Years: The Short Form (SCBE-30)." *Psychological Assessment*, vol. 8, no. 4, 1996, pp. 369-377.

LaFreniere, Peter J., and Jean E. Dumas. *Social Competence and Behavior Evaluation, Preschool Edition*. Los Angeles: Western Psychological Services, 1995.

LaFreniere, Peter J., Jean E. Dumas, France Capuano, and Diane Dubeau. "Development and Validation of the Preschool Socioaffective Profile." *Psychological Assessment*, vol. 4, no. 4, 1992, pp. 442-450.

Madle, Ronald A. "Review of Social Competence and Behavior Evaluation. Preschool Edition." In *The Fourteenth Mental Measurements Yearbook*, edited by Barbara S. Plake and James C. Impara. Lincoln, NE: The Buros Institute of Mental Measurements, 2001.

Poteat, G. Michael. "Review of Social Competence and Behavior Evaluation, Preschool Edition." In *The Fourteenth Mental Measurements Yearbook*, edited by Barbara S. Plake and James C. Impara. Lincoln, NE: The Buros Institute of Mental Measurements, 2001.

SOCIAL SCIENCE READING COMPREHENSION ASSESSMENT, 2007

Authors: Educational Testing Service (ETS)		**Type of Assessment:** Group-administered assessment **Domain:** Reading (comprehension)
Publisher: Educational Testing Service 609-921-9000 http://www.ets.org		**Grade/Age Range:** Grade 5 **Administration Interval:** Annual[1]
Material, Training, and Scoring Costs: Not specified		**Personnel and Training Requirements** **Credentials Required for Use:** Not specified **Personnel for Administration:** Individual with basic clerical skills with some training[1] **Training for Administration:** Minimal (1 to 2 hours)[1]
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: The norming sample consisted of 2,930 students (47 percent male and 47 percent female, and 6 percent unaccounted for) in grade 5. The assessment was administered in 2007. The sample was not nationally representative, but it spanned 89 schools in a geographically diverse area.		**Summary** **Initial Material Cost:** To be determined upon negotiation with the publisher **Time to Administer:** About 45 to 50 minutes[1] **Ease of Administration and Scoring:** 4 (administered or scored by a clinician or specialist) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 2 (older than 10 years or not nationally representative)

NARRATIVE

Description: The Social Science Reading Comprehension Assessment is a group-administered assessment to determine grade 5 students' comprehension of expository text in history or social science. Students read passages and complete multiple-choice items aimed at assessing general reading comprehension skills (for example, vocabulary, main idea identification, inference, purpose) from social science-based passages. Students are not expected to bring a certain level of social science knowledge to the assessment; answers regarding specific social sciences knowledge must be gleaned from the passage. The paper-and-pencil assessment comprises five passages with 30 multiple-choice items in total. It is untimed but is designed to take from 45 to 50 minutes for completion (N. Carey, personal communication, December 16, 2008). The Social Science Reading Comprehension Assessment was developed in particular for the NCEE's Evaluation of Reading Comprehension Programs, which focused on grade 5 students.

Other Languages: None.

Uses of Information: The Social Science Reading Comprehension Assessment may be used to assess grade 5 students' ability to comprehend expository text focused on social science content. In particular, the Evaluation of Reading Comprehension Programs used the measure to evaluate such reading comprehension interventions.

Methods of Scoring: Each item answered correctly receives a score of one point. Total raw scores are the sum of correct items. The developers used Item Response Theory (IRT) to estimate theta scores based on a three-parameter logistic IRT model. The IRT theta scores were then linearly transformed to a scale with a mean of approximately 500 and a standard deviation of 30. ETS conducts the scoring. The Technical Report provides tables to convert the raw score into a scale score.

Interpretability: A higher score means a higher level of ability to comprehend social science-based reading passages. The IRT scale scores are based on the current sample of grade 5 students.

Reliability:
(1) Internal consistency reliability: Cronbach's alpha for raw scores was 0.84 (N = 2,927), calculated from the norming sample but excluding those who responded to fewer than five items.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: No information available.

Validity Evidence:
To determine their appropriateness for inclusion in the assessment, passages and items underwent review by three content experts, an editor (for clarity, spelling, grammar, style), and a researcher (for factual accuracy). In addition, the developers selected passages according to the following criteria: a social science focus (inclusion of at least one passage focused on the history of another country or U.S. history after 1900); readability at a grade 4 and 5 level; and reflective of interests of grade 5 students. The following criteria guided the selection of questions: assessment of general standards of vocabulary, main idea, details, inference, and purpose;

vocabulary and sentence structure appropriate for grade level; and items that form a distribution in difficulty with25 percent of low difficulty, 50 percent medium, and 25 percent high. Experts from both ETS and Mathematica reviewed the passages and items, and the assessment was then piloted,

Construct/Concurrent validity: The developers conducted Classical Item Analysis and flagged items for review on the basis of five characteristics: item difficulty *p*-value less than 0.25 or greater than 0.90; correlation between a correct item response and total test performance less than 0.15; correlation between an incorrect item response and total test performance greater than 0; an item omission rate of 5 percent or greater; or a lack of selection of certain options on an item. Three items were flagged for review by a content expert and a psychometrician. After the three items were reviewed and bias analysis was conducted (see Bias Analysis), it was determined that all items were valid. The developers used all items in IRT calibrations for scoring.

Correlation between students' scores on the Social Science Reading Comprehension Assessment and the Group Reading Assessment and Diagnostic Evaluation (GRADE) in the spring was 0.68. In addition, correlation between class-level means on the Social Science Reading Comprehension Assessment and the Science Reading Comprehension Assessment (a similar assessment using expository text on science topics) was 0.77 for approximately 270 classrooms. (Student-level scores on both the Science and Social Science assessments were unavailable for a given student because students were randomly assigned to take only one of the comprehension assessments. However, student-level scores were available to permit correlation with the GRADE. The class-level correlations might be higher than what would be expected at the student-level given that the class-level correlation with the GRADE was higher than the student-level correlation noted above.)

Predictive validity: No information available.

Bias Analysis: During test development, an ETS staff member trained in issues of fairness reviewed potential items for inclusion, devoting special attention to language use, offensive content, and racial/ethnic or gender bias. After piloting the assessment, the developers conducted differential item functioning (DIF) analyses based on gender and race/ethnicity. Race/ethnicity DIF analyses focused on White, Black, and Hispanic students because sample sizes for other ethnic groups were too small. The analysis found no items exhibiting DIF across subgroups.

Training Support: A two-hour training session reviews the goal of the test and basic proctoring information as well as how to read instructions verbatim and how to respond to students' questions (N. Carey, personal communication, December 16, 2008).

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: None.

Previous Version: None.

NCEE or REL Study Use:[2] Evaluation of Reading Comprehension Programs

[1] N. Carey, personal communication, December 16, 2008.

[2] See Table F.1 for web address.

References:

Educational Testing Service. "Mathematica Reading Comprehension Assessments 2007 Technical Report." Princeton, NJ: Educational Testing Service, 2007a.

Educational Testing Service. *Social Studies Reading Comprehension Assessment* (unpublished). Princeton, NJ: ETS, 2007b.

James-Burdumy, Susanne, David Myers, John Deke, Wendy Mansfield, Russell Gersten, Joseph Dimino, Jan Dole, Lauren Liang, Sharon Vaughn, and Meaghan Edmonds. "The National Evaluation of Reading Comprehension Interventions: Design Report, Final Report." Submitted to the Institute of Education Sciences, U.S. Department of Education. Princeton, NJ: Mathematica Policy Research, 2006.

James-Burdumy, Susanne, Wendy Mansfield, John Deke, Nancy Carey, Julieta Lugo-Gil, Alan Hershey, Aaron Douglas, Russell Gersten, Rebecca Newman-Gonchar, Joseph Dimino, and Bonnie Faddis. "Effectiveness of Selected Supplemental Reading Comprehension Interventions: Impacts on a First Cohort of Fifth-Grade Students." (NCEE 2009-4032). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

SOCIAL SKILLS RATING SYSTEM (SSRS), 1990[1]

Authors: Frank M. Gresham and Stephen N. Elliott		**Type of Assessment:** Student self-report, teacher report, and parent report **Domain:** Social-emotional
Publisher: Pearson Assessments 800-627-7271 www.pearsonassessments.com		**Grade/Age Range:** 3 through 18 years **Administration Interval:** None described
Material, Training, and Scoring Costs: Preschool/Elementary Starter Set (10 copies each of Teacher and Parent Preschool and Elementary Questionnaires and Student Elementary Questionnaire, 10 Assessment Intervention Records, and manual): $147 Preschool/Elementary Starter Set with ASSIST scoring: $385 Secondary Starter Set (10 copies each of Teacher, Parent, and Student Questionnaires, 10 Assessment Intervention Records, and manual): $134 Secondary Starter Set with ASSIST scoring: $365		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2 (bachelor's degree with coursework in measurement and domain) **Personnel for Administration:** Individual with basic clerical skills and some training **Training for Administration:** Self-training (< 1 hour)
Languages: English, Spanish		**Alternate Forms:** No
Representativeness of Norming Sample: Standardization data were collected in 1988 with a nationally representative sample of 4,170 students in grades 3 through 12 (self-ratings for each student and corresponding ratings by 1,027 parents and 259 teachers). The student form sampling plan targeted students in grades 3 through 10; the number per grade varied from 341 to 617. Forty-four grade 11 students and 80 grade 12 students were added. Males and females were equally represented. Students from certain ethnic backgrounds (Black and White), regions (South and North Central), and communities (urban and suburban) were overrepresented. Special education students were oversampled. Separate norms exist for males and females and for elementary students with disabilities (by gender) assessed with the teacher form. Preschool norms are based on 1987 data from a non-representative sample of 200 students.		**Summary** **Initial Material Cost:** 2 ($100 to $200) **Time to Administer:** 15 to 22 minutes for each questionnaire **Ease of Administration and Scoring:** 2 (self-administered or administered and scored by someone with basic clerical skills) **Reliability:** Teacher and Parent Forms: 3 (meets minimum acceptability ratings—0.70) Student Forms: 2 (all or mostly under 0.70) **Predictive Validity:** Not Available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 2 (older than 10 years or not nationally representative)

NARRATIVE

Description: The SSRS is a multirater (student self-report, teacher report, and parent report) measure of the perceived frequency and importance of students' social behaviors for preschool- through secondary school-age students. Teacher report forms (SSRS-T) and parent report forms (SSRS-P) are available for three developmental levels—preschool (age 3 through 4 years), elementary (kindergarten through grade 6), and secondary (grades 7 through 12). Two levels of student self-report forms (SSRS-S) are available (grades 3 through 6 and grades 7 through 12). Respondents complete a paper-and-pencil questionnaire, with the number of items ranging from 34 (SSRS-S Elementary) to 57 (SSRS-T Elementary). Administration times range from approximately 15 to 22 minutes.

The SSRS assesses three general domains represented by corresponding scales: (1) Social Skills, (2) Problem Behaviors, and (3) Academic Competence. The Social Skills domain comprises five subscales (Cooperation, Assertion, Self-Control, Responsibility, and Empathy). The SSRS-T, SSRS-P, and SSRS-S include Cooperation, Assertion, and Self-Control subscales; the SSRS-P includes a Responsibility subscale, and the SSRS-S includes an Empathy subscale. Respondents rate the frequency of behaviors (0 = never, 1 = sometimes, 2 = very often); for all forms except the SSRS-S Elementary, they also rate the perceived importance of a behavior. The Problem Behaviors domain is assessed with items forming three subscales (Externalizing Problems, Internalizing Problems, and Hyperactivity). Problem Behaviors items are included on the SSRS-T and SSRS-P only; respondents provide frequency but not importance ratings. Finally, the Academic Competence domain is measured by items assessing mathematics and reading performance, motivation, parental support, and cognitive ability. The items appear in the SSRS-T Elementary and SSRS-T Secondary only.

Other Languages: Jurado et al. (2006) investigated the reliability, validity, and cultural adequacy of a Spanish-language version of the Social Skills scale of the SSRS-T Elementary. They administered it to 44 teachers of 357 grade 1, 3, and 5 students in urban and rural Puerto Rico. The authors reported alpha coefficients of 0.95 for the total scale score and 0.89 to 0.92 for the subscale scores. Test-retest reliability coefficients were 0.87 for the total scale and ranged from 0.69 to 0.92 for the subscales. The authors reported a correlation of -0.69 between scores on the SSRS-T Social Skills scale and the Problem Behaviors scale on the Teacher Report Form (TRF) by Achenbach. Correlations between the SSRS-T Social Skills subscales (Cooperation, Assertion, and Self-Control) and TRF Problem Behaviors subscales (Internalizing and Externalizing) ranged from -0.20 to -0.64. Scores on the SSRS-T Social Skills scale also varied by age and gender; younger students and females received higher social skills ratings from teachers. The authors caution that sample limitations constrain the generalizability of their findings. Use of the Spanish adaptation requires special permission from the SSRS publisher.

Uses of Information: The SSRS may be used to assess students' social behaviors for both research and clinical applications. It may be used to assess individual students or groups, and its multirater format enables the collection of information about a student's social functioning from several sources. The SSRS is widely used to identify students who display significant problem behaviors and to inform educational decision making, placement, and intervention planning for identified students.

Methods of Scoring: For the SSRS-T and SSRS-P, the response format for Social Skills and Problem Behaviors items involves both frequency and importance ratings for each item along a 3-point scale. The SSRS-S assesses frequencies of social skills behaviors along the same 3-point scale. Assessors determine total subscale raw scores (the sum of the raw frequency scores for all items on each subscale) and convert them into behavioral categories called Behavior Levels (labeled Fewer, Average, or More based on whether they fall below, within, or above 1 standard deviation from the mean, respectively). They also determine Social Skills, Problem Behaviors, and Academic Competence scale scores by summing the total subscale raw scores across the subscales comprising each scale. Total raw scores for each scale are converted to Behavior Levels, standard scores (with confidence bands), and percentile ranks. Importance ratings are generally used for intervention planning.

Separate scoring procedures are required for Academic Competence items on the SSRS-T forms. The response format is a 5-point scale on which teachers rate a student's percentile rank on an academic characteristic (e.g., 1 = lowest 10 percent, 5 = highest 10 percent).

Interpretability: User qualifications require SSRS results to be interpreted by individuals with at least a bachelor's degree who have completed coursework in testing and measurement or who otherwise have permission to administer such an assessment in their jurisdiction. SSRS interpretation involves comparing a student's scores to peer group scores (interindividual comparisons) and analysis of individual score patterns across subscales and scales (intra-individual comparisons). Intra-individual comparisons of scores (indicating individual strengths and weaknesses), item-level interpretation, and importance ratings may be useful for intervention planning.

Reliability:
(1) Internal consistency reliability: Across all eight SSRS forms, Cronbach's alpha coefficients for the scores from the Social Skills scale ranged from 0.83 to 0.94. For the Problem Behaviors scale, the six SSRS-T and SSRS-P forms for each developmental level yielded alpha coefficients ranging from 0.77 to 0.88. The alpha coefficient for scores from the Academic Competence scale (assessed on the SSRS-T Elementary and Secondary) was 0.95. Internal consistency estimates were lower for scores from subscales and varied across the SSRS-T, SSRS-P, and SSRS-S. Alpha coefficients for scores from subscales ranged from 0.74 to 0.92 for the SSRS-T, 0.57 to 0.83 for the SSRS-P (17 of 19 were above 0.70), and 0.51 to 0.77 for the SSRS-S (2 of 8 were above 0.70).
(2) Test-retest reliability: Samples of teachers, parents, and students from the elementary level standardization sample completed a second assessment four weeks after their initial ratings. For the overall scales, test-retest correlation coefficients for teacher ratings were 0.85, 0.84, and 0.93 for Social Skills, Problem Behaviors, and Academic Competence, respectively. Parent test-retest reliability coefficients were 0.87 for Social Skills and 0.65 for Problem Behaviors; for student self-ratings, the test-retest reliability coefficient was 0.68. For the subscales, Social Skills yielded test-retest reliability coefficients ranging from 0.75 to 0.88 for teachers, 0.77 to 0.84 for parents, and 0.52 to 0.66 for students. Problem Behavior scores demonstrated test-retest reliability coefficients ranging from 0.76 to 0.82 for teachers and from 0.48 to 0.72 for parents.
(3) Alternate form reliability: No alternate forms.

(4) Inter-rater reliability: No information available using the different raters within the same context (e.g., a mother and father completing the SSRS-P). See Construct/Convergent validity for correlations between raters across contexts and forms.

Validity Evidence:
SSRS items were developed in accordance with empirical research published through the 1980s on social skills assessment and training, relationships between social behaviors and social outcomes in children and youth, and differences in social behaviors between students with and without disabilities.

Construct/Concurrent validity: Construct validity was investigated through eight sets of factor analyses—one for each domain of the SSRS-T, SSRS-P, and SSRS-S across developmental levels. Separate analyses of the SSRS-T (with preschool, elementary, and secondary samples of 212, 1,033, and 318 students, respectively) yielded Cooperation, Assertion, and Self-Control factors as well as an Academic Competence factor in the elementary and secondary samples. Problem Behaviors items formed Externalizing and Internalizing factors as well as a Hyperactivity factor in the elementary sample only. Analyses of parent ratings found Cooperation, Assertion, Responsibility, and Self-Control social skills factors; Externalizing and Internalizing problem behavior factors; and, in the elementary data only, a Hyperactivity problem behavior factor. Analyses of student self-ratings of 2,407 elementary students in the standardization sample yielded four social skills factors (Cooperation, Assertion, Responsibility, and Self-Control) and two behavior problems factors (Internalizing and Externalizing). In analyses of the SSRS-S Secondary (conducted with 1,770 standardization sample students), four social skills factors emerged: Cooperation, Assertion, Self-Control, and Empathy.

The SSRS manual reports correlations ranging from 0.16 to 0.25 between parent and teacher ratings on Social Skills items on the SSRS-P Preschool and SSRS-T Preschool forms. Elementary-level convergent validity correlation coefficients between teacher, parent, and student ratings ranged from 0.03 to 0.41 (14 of 16 were statistically significant at the 0.05 level or greater). Correlations among teacher, parent, and student ratings on the secondary-level SSRS ranged from 0.19 to 0.43 on the Social Skills scale and subscales and from 0.10 to 0.22 on the Problem Behaviors scale and subscales.

The SSRS manual also reports average convergent validity (Fisher's *z* transformation) of scores across raters. Across preschool, elementary, and secondary levels, the average Social Skills coefficients were 0.31 for teacher and parent ratings, 0.32 for teacher and student ratings, and 0.24 for parent and student ratings. For Problem Behaviors, the convergent validity coefficient was 0.29 between teacher and parent ratings.

For the SSRS-T, three validity studies of standardization sample data demonstrated correlations between SSRS-T ratings and other social skills measures. The first study compared teacher ratings of 79 elementary students on the SSRS-T and the Social Behavior Assessment (SBA), a measure on which high scores indicate behavior problems. The authors reported a correlation of 0.55 between total scores on the SBA and the SSRS-T Problem Behaviors scale. They also reported correlations ranging from 0.01 to 0.57 between SBA subscales and SSRS-T Problem Behaviors subscales.

The second study compared teacher ratings of 99 elementary students on the SSRS and the Child Behavior Checklist-Teacher Report Form (CBCL-TRF). The authors reported correlations of 0.75 between SSRS and CBLS Externalizing subscale scores and of 0.59 between the measures' Internalizing subscale scores. The third study compared ratings of 269 students on the SSRS-T Elementary and the Harter Teacher Rating Scale (TRS). Correlations between the TRS total score and the SSRS-T Social Skills and Academic Competence scales were 0.70 and 0.63, respectively.

The authors also correlated scores on the SSRS-P Elementary and the Child Behavior Checklist-Parent Report Form (CBCL-PRF) for 46 students from the standardization sample. The correlation between the SSRS-P Problem Behaviors scale and the CBCL-PRF was 0.70; between the measures' corresponding total Social Skills scales, it was 0.58. Correlations between corresponding Internalizing and Externalizing subscales ranged from 0.42 to 0.70.

Two studies showed that scores on the SSRS-S Elementary demonstrated lower correlations than the Teacher and Parent forms with scores on criterion measures (Gresham and Elliott 1990). The first study (conducted with 42 students from the standardization sample) compared scores on the SSRS-S Elementary and Child Behavior Checklist-Youth Self-Report Form (CBCL-YSR), a measure of behavior problems and social competence. The correlation between scores from the SSRS-S Social Skills scale and CBCL-YSR Social Competence scale was 0.23. The authors reported that most correlations between the SSRS-S Social Skills subscales and CBCL-YSR Social Competence subscales were nonsignificant and near zero, but they offered no explanation for their findings. The second study compared scores on the SSRS-S Social Skills scale and subscales and the Piers-Harris Children's Self-Concept Scale (PHCSCS) with 79 students from the standardization sample. Many of the correlations were nonsignificant, except between SSRS-S Social Skills subscales and PHCSCS scores for Behavior and Intellectual and School Status (correlations ranged from 0.18 to 0.41).

For the SSRS-T, the authors reported correlations of -0.67 and -0.68 between the SSRS-T Academic Competence and Social Skills scales, respectively, and the SBA total score. Correlations between SSRS-T Social Skills subscales and SBA subscales ranged from -0.15 to -0.73. The authors reported a correlation of -0.66 between the SSRS-T Problem Behaviors scale and the TRS total score. Correlations between the SSRS-T Externalizing, Internalizing, and Hyperactivity subscales and the TRS total score were -0.50, -0.44, and -0.57, respectively.

For the SSRS-P Elementary, the Social Skills scale correlated -0.37 with the CBCL-PRF Behavior Problems scale. Correlations between the SSRS-P Social Skills subscales and CBCL-PRF subscales ranged from -0.11 to -0.43. The SSRS-P Problem Behaviors scale correlated -0.52 with the CBCL-PRF Social Competence scale. The authors reported correlations ranging from -0.03 to -0.61 on the SSRS-P Problem Behaviors subscales and two of the CBCL-PRF Social Competence subscales (Social Functioning and School Functioning). They also reported near-zero correlations between the SSRS-P Problem Behaviors subscales and the CBCL-PRF Activities subscale.

Correlations between the SSRS-S Social Skills scale and CBCL-YSR Behavior Problems scales were 0.23 and -0.33, respectively. The authors reported correlations ranging from -0.21 to -0.48 between the SSRS-S Social Skills subscales and CBCL-YSR Externalizing subscales. Only one

SSRS-S Social Skills subscale, Cooperation, correlated significantly with the Internalizing subscale on the CBCL-YSR (r = -0.27). The authors reported that most correlations between the SSRS-S Social Skills subscales and CBCL-YSR Social Competence subscales were nonsignificant and near zero.

The SSRS manual also reports discriminant (or divergent) validity coefficients comparing student and parent ratings, student and teacher ratings, and teacher and parent ratings on the Social Skills total scale and subscales. Across forms (Preschool, Elementary, and Secondary), the coefficients ranged from -0.01 to 0.39 (student-parent), 0.06 to 0.34 (student-teacher), and 0.04 to 0.28 (teacher-parent). The authors point to these low correlations between different subscales measured by different raters as evidence of discriminant validity. They note that higher correlations between different Social Skills subscales assessed by the same raters provide less evidence of discriminant validity (median r = 0.54, 0.48, and 0.50 for teacher, parent, and student ratings, respectively).

Teacher, parent, and student rating scores from the standardization sample on the Social Skills, Problem Behaviors, and Academic Competence scales did not vary significantly by age/grade but did exhibit significant gender differences. Social skills ratings for students at almost every grade level from preschool through grade 10—as rated by teachers, parents, and students themselves—were higher for females than for males. Conversely, teachers and parents consistently rated males as exhibiting more frequent problem behaviors than females. Separate norms are available by gender.

The SSRS manual also reports that, across developmental levels and informant types, students without disabilities were rated higher in social skills than students with disabilities. Significantly different scores were also observed between learning-disabled and other types of disabled students on particular SSRS components.

Predictive validity: No information available.

Bias Analysis: Researchers have investigated whether the Social Skills scale and subscales of the SSRS-T Preschool and Elementary forms exhibit construct bias when used with samples of minority students. With a sample of 943 predominantly Black, urban Head Start children, Fantuzzo et al. (1998) found three social skills factors—Self-Control, Assertion, and Interpersonal Skills (the Cooperation factor was not replicated in this sample). The Self-Control, Interpersonal Skills, and Cooperation factors were replicated, however, in a study with an ethnically diverse, national sample of 958 typically developing grade 1 students (Walthall et al. 2005). The factor structures were invariant for White and minority subgroups.

Training Support: No information available.

Adaptations/Special Instructions for Individuals with Disabilities: Questionnaire items may be read to respondents who are unable to read.

Alternate Forms: None.

Previous Version: None.

NCEE or REL Study Use:[2] Lessons in Character Education (REL-West)

[1] To be included in the compendium, information sources for a measure had to be available by the review team by mid-November 2008. A revised version of the SSRS—Social Skills Improvement System (SSIS)—was released after this date in late 2008.

[2] See Table F.1 for web address.

References:

Fantuzzo, John, Patricia Manz, and Paul McDermott. "Preschool Version of the Social Skills Rating System: An Empirical Analysis of its Use with Low-Income Children." *Journal of School Psychology*, vol. 36, no. 2, 1998, pp. 199-214.

Gresham, Frank M., and Stephen N. Elliott. *Social Skills Rating System*. Bloomington, MN: Pearson Assessments, 1990.

Gresham, Frank M., and Stephen N. Elliott. *Social Skills Rating System Manual*. Circle Pines, MN: American Guidance Service, 1990.

Jurado, Michelle, Eduardo Cumba-Aviles, Luis Collazo, and Maribel Matos. "Reliability and Validity of a Spanish Version of the Social Skills Rating System--Teacher Form." *Journal of Psychoeducational Assessment*, vol. 24, no. 3, 2006, pp. 195-209.

Walthall, Johanna, Timothy R. Konold, and Robert C. Pianta. "Factor Structure of the Social Skills Rating System Across Child Gender and Ethnicity." *Journal of Psychoeducational Assessment,* vol. 23, 2005, pp. 201-215.

STANFORD ACHIEVEMENT TEST SERIES, TENTH EDITION (STANFORD 10), 2003

Authors: Harcourt Assessment, Inc.		**Type of Assessment:** Group-administered assessment **Domain:** Reading (phonological awareness, letter recognition and naming, vocabulary, print concepts, decoding, phonics, comprehension skills), language arts/language proficiency (writing, editing skills, grammar, spelling, conventions, syntax, vocabulary, morphology, listening skills), mathematics, science, and social studies
Publisher: Pearson Education, Inc. 800-211-8378 http://www.pearsonassess.com		**Grade/Age Range:** Kindergarten through grade 12 **Administration Interval:** None described
Material, Training, and Scoring Costs: Test Kit (one copy of complete battery, test booklet, administration directions, practice test, answer document): $52 per level, form 10 machine-scorable test booklets (Primary 1 to Primary 3, sold separately): $88.35 10 reusable test booklets (Intermediate through TASK, sold separately): $74.25 30 machine-scorable answer sheets: $52.75 per level, form Technical Data Report: $52.00 Multilevel Norms Books: $69.30 (separate books for fall and spring norms)		**Personnel and Training Requirements** **Credentials Required for Use:** Not specified, except that test is sold only to schools and school districts **Personnel for Administration:** Individual with basic clerical skills with some training **Training for Administration:** Minimal (1 to 2 hours)
Languages: English		**Alternate Forms:** Two sets of equivalent forms (A and B; D and E) for levels Primary 1 through TASK
Representativeness of Norming Sample: Data were collected with spring and fall 2002 samples of 250,000 and 110,000 students, respectively. The developers used stratified cluster (i.e., classroom) sampling, repeated within each of 48 states so that within-state samples would be similar to the national population. The sample (with test scores weighted) approximated levels in the 2000 Census and 2000–2001 National Center for Education Statistics data for region, socioeconomic status, urbanicity, and ethnicity and included students with disabilities in regular education classrooms.		**Summary** **Initial Material Cost:** 2 ($100 to $200) **Time to Administer:** Approximately 2 to 5 hours depending on test (see Description) **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3[1] (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 3[2] (normed within past 10 years and nationally representative)

NARRATIVE

Description: The Stanford Achievement Test Series Tenth Edition (Stanford 10) is a group-administered, multiple-choice assessment of students' school achievement in reading, mathematics, spelling, language, science, social science, and listening. It is a battery of 13 test levels assessing students from kindergarten through grade 12. The Stanford Early School Achievement Test (SESAT) consists of two test levels for assessing students in kindergarten and the first half of grade 1. The SESAT 1 is used with students from the beginning to the middle of kindergarten; the SESAT 2 is used with students in the middle of kindergarten through the middle of grade 1. The Stanford Achievement Test consists of eight levels for assessing students from the second half of grade 1 through the end of grade 9: Primary 1 (grades 1.5 to 2.5), Primary 2 (grades 2.5 to 3.5), Primary 3 (grades 3.5 to 4.5), Intermediate 1 (grades 4.5 to 5.5), Intermediate 2 (grades 5.5 to 6.5), Intermediate 3 (grades 6.5 to 7.5), Advanced 1 (grades 7.5 to 8.5), and Advanced 2 (grades 8.5 to 9.9). Finally, the Stanford Test of Academic Skills (TASK) consists of three levels for assessing basic skills in grades 9 through 12.

At all test levels, assessors may administer the full-length Complete Battery or the Abbreviated Battery. The SESAT 1 and 2 levels have only one form. All other test levels have two versions (each with alternate forms) that assess the same subtest areas except for the language subtest they include. One version (Forms A and B) includes the Traditional Language subtest, which focuses on mechanics and expression; the other version (Forms D and E) includes the Comprehensive Language subtest, which focuses on writing processes such as prewriting, composing, and editing.

Stanford 10 test levels include different combinations of subtests. All test levels have subtests in Reading and Mathematics, the number of which varies by level. The Primary 1 through TASK 3 levels have Language subtests, and the Primary 3 and higher levels also have a Science and Social Studies subtest. In the lower levels, the Environment subtest assesses student achievement in science and social studies. The SESAT 1 through Advanced 2 levels also have a Listening subtest.

All Stanford 10 items are written in multiple-choice format. Some items include additional open-ended questions and writing prompts. The developers designed each item to assess four achievement parameters: (1) a content cluster, (2) a process cluster, (3) a cognitive level (basic thinking or thinking skills), and (4) an instructional standard. The number of items in each subtest of the Complete Battery varies from 30 to 54; in the Abbreviated Battery, subtests consist of either 20 or 30 items. To reduce student frustration and anxiety, items are arranged in a format that mixes easy and difficult items rather than in an easy-to-hard order.

The assessment may be administered in paper-and-pencil format (with hand- or machine-scorable answer sheets) or online through the Pearson web site. The assessment is untimed, although the administration instructions include estimated time allocations for each subtest. Administration of the Complete Battery requires approximately 2.5 hours for the SESAT; 5.25 hours for the Primary, Intermediate, and Advanced levels; and 3.75 hours for the TASK levels. Assessors may administer practice tests for each test level within a week of actual testing.

The Technical Data Report notes a ceiling effect for the Sentence Reading subtest at the Primary 1 test level. Many grade 1 students earn perfect scores on this subtest, perhaps rendering the

subtest unsuitable for evaluating higher-achieving students. For such students, the developers recommend the total Reading score or the total Reading score computed without the Sentence Reading subtest score.

Other Languages: None.

Uses of Information: Educators, evaluators, or researchers may use Stanford 10 scores to compare student achievement to that of a nationally representative group of students. In addition, the vertically linked scale scores permit the longitudinal tracking of students' progress. Assessment results may be examined in terms of specific skills (subtest scores) or general skill areas (cluster or composite scores).

Methods of Scoring: The Stanford 10 may be hand- or machine-scored locally or sent to the publisher for scoring and reporting. The Stanford 10 reports results in terms of raw scores, scaled scores, individual percentile ranks, stanines, grade equivalents, normal curve equivalents (NCE), achievement/ability comparisons (AAC) between scores on the Stanford 10 and the Otis-Lennon School Ability Test Eighth Edition (OLSAT 8), group percentile ranks and stanines, content cluster and process cluster performance categories, item *p*-values (to allow comparisons of item difficulties for a local group versus the normative sample), and criterion-referenced performance levels (below basic, basic, proficient, advanced). Assessors using the Complete Battery may use individual and group score reports and summaries to present and interpret results.

Beginning at the Primary 3 level for both the Complete and Abbreviated batteries, assessors may calculate a Thinking Skills score by summing the number of correct responses across relevant subtests and using norm tables to convert the raw score into a scaled score.

Interpretability: The Technical Data Report states that scaled scores express performance across all test levels of any subtest on a single scale but are not comparable across content areas or subtests within a content domain. During development of the Stanford 10, the developers formed subsamples of students in the norming sample to allow for linking of data across Stanford 10 test levels, forms, and editions. In doing so, they (1) created a continuous, vertical scale that allows for comparison of scores across levels of the battery and (2) established the equivalence of scores on the test's alternate forms (see Alternate form reliability) and on the Stanford 10 and Stanford 9 (see Construct/Concurrent validity). While vertical scaling across test levels allows for cross-grade comparisons of scaled scores, Morse (2005) cautioned users that ". . .median year-to-year increases in scaled scores diminish in size with increasing grade level and can, with some subtests of the TASK, show little or no increase or even a decrease" (p. 975).

Reliability:
(1) Internal consistency reliability: The Technical Data Report provides Kuder-Richardson Formula 20 (KR20) coefficients as estimates of internal consistency reliability. For fall and spring scores on all forms of the full-length SESAT 2 through TASK tests, KR20 coefficients ranged from 0.86 to 0.97 for composite scores and from 0.54 to 0.97 for subtest scores, with the majority of coefficients in the mid- 0.80s to 0.90s. Only nine of the 409 KR20 coefficients for subtest scores were below 0.70, and all of them corresponded to scores from the Prewriting and Composing subtests of the Comprehensive Language domain (Forms D and E). At all test levels, KR20 coefficients for scores from the Abbreviated Battery tended to be lower than those from the Complete Battery (coefficients for most subtest and total test scores on the Abbreviated

Battery were in the 0.70s and 0.80s). Exceptions included scores from the Environment subtest, Comprehensive Language subtests, and Social Science subtest, all of which had several corresponding KR20 coefficients below 0.70.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: The Technical Data Report provides alternate form reliability coefficients, correlating scores on Forms A and B with scores on Forms D and E. The developers conducted alternate form reliability analyses with data from subsamples of students in the standardization sample who completed two forms of the test (sample sizes ranged from 144 to 915). Correlations between corresponding subtest scores from Forms A and B of the Primary 1 through TASK 3 levels ranged from 0.63 to 0.93 (the majority of coefficients were above 0.70). For Reading, Mathematics, and Language composite scores, coefficients were higher, ranging from 0.72 to 0.93. Correlations between corresponding subtest scores from Forms D and E ranged from 0.53 to 0.78; between composite scores, correlations ranged from 0.74 to 0.89.
(4) Inter-rater reliability: No information available.

Validity Evidence:
In developing the Stanford 10 items, developers reviewed national and state instructional standards, state and school district content-specific curricula, major textbook series in every subject area, and current educational trends identified by national professional educational groups. The developers designed the items to align with standards-based national curricula and to assess concepts and skills normally taught during the second half of a given school year and the first half of the next year. The Stanford 10 developers recommend that users examine the alignment of the content to their own district's goals and curricula.

Developers administered a preliminary version of the Stanford 10 to a nationally representative item tryout sample of 170,000 students in 1998 through 2000. They used classical item-analysis methods, such as examining item *p*-values, item correlations with subtest total scores (subtest median biserial correlation coefficients ranged from 0.37 to 1.00), and item discrimination (subtest median point-biserial coefficients ranged from 0.27 to 0.57). In addition, the Technical Data Report states that the developers used Rasch model techniques to calibrate scale scores and estimate item characteristics, such as item difficulty estimates and mean-square fit (values for which are not provided). The developers first selected items for the abbreviated forms and then added items to this core group to form the full-length forms. Given that the abbreviated forms are "core subsets" of the items on the full-length forms, the developers state that validity information for full-length forms applies equally to abbreviated forms (except for lower reliabilities for scores from abbreviated forms owing to the smaller number of items).

Construct/Concurrent validity: The Technical Data Report provides mean scaled scores on each level and subtest for the students in the 2002 norming sample who demonstrate growth over time.

In addition, the Stanford 10 developers reported Pearson product-moment correlation coefficients between corresponding subtest and total test scores on the Stanford 9 and Stanford 10. Across test levels, coefficients ranged from 0.46 to 0.92, with most in the 0.70s and 0.80s. The developers also reported correlations between scores on the Stanford 10 Forms A and D with scores from the OLSAT 8. Intercorrelations between Stanford 10 composite scores and OLSAT 8 verbal, nonverbal, and total scores ranged from 0.35 to 0.83, with the majority of

coefficients in the 0.40s to 0.60s. Similar intercorrelations were reported between Stanford 10 subtest scores and OLSAT 9 verbal, nonverbal, and total scores.

To create a continuous vertical scale permitting interpretation of scores across test levels, the Stanford 10 developers administered two adjacent test levels (e.g., SESAT 1 and SESAT 2, SESAT 2 and Primary 1, and so forth) of each subtest to samples of students ranging in size from 135 to 1,511. Students completed tests at their own grade level and one grade level lower. Intercorrelations between students' corresponding Stanford 10 subtest and total scores on adjacent levels of the test ranged from 0.47 to 0.93 (Harcourt Assessment, Inc. 2004).

Predictive validity: No information available.

Bias Analysis: Items underwent review by assessment specialists or item writers and members of a bias review panel; the panel consisted of 20 educational experts who reviewed items for potential bias with respect to gender, race/ethnicity, religion, geographic region, socioeconomic status, English proficiency, and disability. Experts had diverse racial/ethnic backgrounds and represented a variety of geographic regions and settings (urban, suburban, rural). The panel also included experts in disability issues. Items deemed problematic by the panel were removed from the item pool.

Developers also analyzed all national item tryout program items by using Mantel-Haenszel procedures to detect differential item functioning (DIF) between majority and minority groups that were matched on test scores. The developers compared item scores of males and females, White and Black students, White and Hispanic students, and students with and without disabilities. The developers reviewed items identified with potential DIF and potentially excluded them from the final forms of the tests.

For items with evidence of DIF that were retained, the developers counterbalanced items that potentially favored one group over another with items that favored the second group over the first. In the test materials, the developers balanced the frequency and types of depictions of minority or gender group members.

Training Support: The administration instructions state that, before test administration, assessors should familiarize themselves with the test level and form they are administering by taking the test themselves and reviewing the directions for administration. No formal training is required.

Adaptations/Special Instructions for Individuals with Disabilities: Assessors may administer Braille and large-print editions to visually impaired students. Separate norms are available for the Braille editions of the assessment. For hearing-impaired students, schools may use screening tests to identify the proper Stanford 10 test level to be administered and then use special norms provided by the publisher to interpret results. The administration instructions specify other allowable accommodations.

Alternate Forms: Forms A and D are identical except that Form A includes the Traditional Language subtest and Form D includes the Comprehensive Language subtest. The Technical Data Report states that Forms A and D are equivalent in content and difficulty to Forms B and E, respectively (see Alternate form reliability). Form A may be used at all test levels, whereas

Forms D, B, and E may be used from the Primary 1 through TASK 3 levels. Complete and Abbreviated batteries are available for all forms.

Previous Version: The Stanford 10 replaces the Stanford 9, which was published in 1996. It updates norms, item content (reflecting current educational standards and curricula), and test materials (including realistic, full-color illustrations like those in textbooks). It consists of all new items for all levels of the assessment. Improvements include easier navigation through multiple-choice test booklets and answer sheets, simplified reports, and reading selections written specifically for the test by children's book authors. Stanford 10 developers equated the test with the Stanford 9 by using data from approximately 1,000 students per test level who were administered both the Stanford 9 and Stanford 10.

NCEE or REL Study Use:[3] Reading First Impact Study; The Evaluation of Enhanced Academic Instruction in After-School Programs; The Effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI) (REL-Southeast); Evaluation of Principles-Based Professional Development to Improve Reading Comprehension for English Language Learners (REL-Pacific); Evaluation of the DC Opportunity Scholarship Program; Evaluation of the Effectiveness of Educational Technology Interventions

[1] This rating refers to reliability coefficients for the total test scores; the individual subtests/subscales encompassed some reliability coefficients below the 0.70 level.

[2]Where sample characteristics differed from national school population characteristics on demographic variables related to test performance, test scores were weighted for better approximation of national characteristics.

[3] See Table F.1 for web address.

References:

Carney, Russell N. "Review of the Stanford Achievement Test, Tenth Edition." In *The Sixteenth Mental Measurements Yearbook*, edited by Robert Spies and Barbara S. Plake. Lincoln, NE: Buros Institute of Mental Measurements, 2005.

Harcourt Assessment, Inc. *Stanford Achievement Test Series (10th Edition). Technical Data Report*. San Antonio, TX: Harcourt Assessment, Inc., 2004.

Harcourt Brace. *Stanford Achievement Test (10th Edition)*. San Antonio, TX: Harcourt Assessment, Inc., 2003.

Harcourt Educational Measurement. *Stanford Achievement Test Series—Ninth Edition*. San Antonio, TX: Harcourt Educational Measurement, 1996.

Morse, David T. "Review of the Stanford Achievement Test, Tenth Edition." In *The Sixteenth Mental Measurements Yearbook*, edited by Robert A. Spies and Barbara S. Plake. Lincoln, NE: Buros Institute of Mental Measurements, 2005.

STANFORD DIAGNOSTIC READING TEST, FOURTH EDITION (SDRT 4), 1995, 2004

Authors: Bjorn Karlsen and Eric F. Gardner		**Type of Assessment:** Group-administered assessment **Domain:** Reading (phonological awareness, letter recognition and naming, print concepts, vocabulary, decoding, phonics, reading fluency, comprehension)
Publisher: Pearson Education, Inc. 800-211-8378 http://www.pearsonassess.com		**Grade/Age Range:** Kindergarten through grade 13 (first semester of college) **Administration Interval:** None described
Material, Training, and Scoring Costs: Red through Blue Level Examination Kit (Test Booklet and Directions; Practice Test and Directions; Answer Document; Class Record Form; Reading Questionnaire, Reading Strategies Survey, Story Retelling Story and Response Form, directions): $61 Pink and Teal Level Examination Kit (Machine-Scorable Test Booklet and Directions, Practice Test and Directions, Flashcards, and Class Record Form): $61 Red, Orange, Green, Purple, or Brown Level Practice Tests (25): $29.30 Pink and Teal Practice Tests (25): $24 Various Test Booklets (Reusable, Hand-Scored, or Machine-Scored, 25): $122.00–$191.55		**Personnel and Training Requirements** **Credentials Required for Use:** 2+ (certification beyond bachelor's like a master's) **Personnel for Administration:** Individual with basic clerical skills with some training **Training for Administration:** Self-training (< 1 hour)
Languages: English		**Alternate Forms:** Two equivalent forms, J and K, for Purple, Brown, and Blue levels
Representativeness of Norming Sample: Normative data were collected in fall 1994 (N = 33,000 kindergarten through grade 12 students and 2,000 college freshmen) and spring 1995 (N = 20,000 kindergarten through grade 12 students) with stratified random samples approximating the U.S. school population according to region, socioeconomic status, ethnicity, and urbanicity. Four percent of students were learning disabled. For the Pink and Teal forms (Kindergarten through first grade), data were collected with stratified (by ethnicity) random samples in spring (N = 3,000) and fall 2004 (N = 4,000). Students with disabilities comprised 7.6 percent of the samples.		**Summary** **Initial Material Cost:** 2 ($100 to $200) **Time to Administer:** 85 to 120 minutes **Ease of Administration and Scoring:** 3 (administered and scored by highly trained individual) **Reliability:** 3 (all at or above 0.70)[1] **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 3 for Pink and Teal levels (normed within past 10 years and nationally representative); 2 for Red through Blue levels (older than 10 years)

NARRATIVE

Description: The Stanford Diagnostic Reading Test (SDRT 4) is a group-administered, paper-and-pencil assessment of strengths and weaknesses in student's reading skills. It may be used to assess students in kindergarten through the first semester of college, and provides both norm-referenced and criterion-referenced information about students' performance. The SDRT 4 assesses reading skills according to a developmental sequence, emphasizing different skills at eight developmental levels: (1) Pink (Grades K.0 to K.5); (2) Teal (Grades K.5 to 1.5); (3) Red (Grades 1.5 to 2.5); (4) Orange (Grades 2.5 to 3.5); (5) Green (Grades 3.5 to 4.5); (6) Purple (Grades 4.5 to 6.5); (7) Brown (Grades 6.5 to 8.9); and (8) Blue (Grades 9.0 to 13.0/first semester of college). The Red through Blue levels were published as the SDRT 4 in 1995; the Pink and Teal levels were added in 2004 to extend use of the test to students in kindergarten and the first half of first grade. The Pink and Teal levels have five subtests assessing essential components of reading instruction as specified by Reading First—Words (Vocabulary), Sounds (Phonemic Awareness), Letters (Phonics Skills), Stories (Comprehension), and Pictures (Fluency). The other test levels have four subtests—Phonetic Analysis, Vocabulary, Comprehension, and either Phonetic Analysis (Red, Orange and Green levels) or Scanning (the ability to scan material quickly for key information; Purple, Brown, and Blue levels). Each subtest measures proficiency in specific skill area objectives that are grouped into clusters and subclusters. For the Pink and Teal levels, the developers also published abbreviated screening tests that identify reading performance at one of three levels. The SDRT 4 also includes three optional informal assessments for which no validity information is provided: (1) the Reading Questionnaire (which assesses reading-related attitudes, habits, and interests); (2) the Reading Strategies Survey; and (3) the Story Retelling subtest (which assesses comprehension through a student's written or oral reconstruction of a story).

Other Languages: None.

Uses of Information: As a diagnostic assessment, the SDRT 4 is designed primarily for use with low-achieving students to identify individual areas of weaknesses and strengths and to inform appropriate instruction. It can also be used to quantify the level of reading performance for groups of students (e.g., classrooms, grades, schools, and school districts). Developers state that educational researchers can use the SDRT 4 to assess the effectiveness of instructional programs or interventions and to measure changes in reading performance over time.

Methods of Scoring: SDRT 4 multiple-choice tests may be hand-scored on site or scored by Pearson Scoring and Reporting Services. For hand-scoring, the technical manual includes a reproducible scoring sheet. Raw scores (the number of correct responses) are converted to scaled scores and progress indicators. Cut scores for progress indicators are specified for each cluster and subcluster to set competence levels needed to make satisfactory progress in a grade-level reading curriculum. Scaled scores may be converted to percentile ranks, stanines, normal curve equivalents (NCE), and grade equivalents. The SDRT 4 scores for students in grades 2 through 12 may be converted to Lexile reading scores. The respective manuals for each level include instructions for scoring the optional measures.

Interpretability: A criterion-referenced interpretation of scores (raw scores and progress indicators) provides information about performance on specific types of test questions and may provide supplemental information about strengths and weaknesses in specific areas. One

reviewer notes, however, that no validity evidence is provided in support of the use of the progress indicators (Engelhard 1998). The authors state that scaled scores are equivalent across alternate forms and comparable across test levels. One reviewer cautions that, across test levels, reliability estimates of scores for several subtests fall below acceptable levels (< 0.70); therefore, the diagnosis of reading difficulties based on subtest scores is not recommended (Engelhard 1998). Publisher scoring options allow test users to order Individual Diagnostic Reports (containing raw, criterion- and norm-referenced scores), Skills Analysis (the number of questions answered correctly out of the number possible for each skill compared to the progress indicator cutoff score), Class Summary Reports, and School and District Summary Reports. The Lexile reading scores for students in grades 2 through 12 may assist in matching students to appropriately challenging reading materials.

Reliability:
(1) Internal consistency reliability: The SDRT 4 manual presents Kuder-Richardson Formula #20 (KR20) and #21 (KR21) reliability coefficients based on the Spring 1995 data (Red through Blue forms). Across test levels and parallel forms, KR20 reliability coefficients for total test scores ranged from 0.95 to 0.97, and coefficients for subtest scores ranged from 0.79 to 0.94. Across test levels and forms, KR21 coefficients for subtest scores ranged from 0.76 to 0.93. The authors reported several reliability coefficients for scores from clusters and subclusters below 0.70.

A separate manual reports KR20 and KR21 coefficients based on data from the 2004 Pink and Teal fall and spring samples. Across samples, KR20 coefficients ranged from 0.90 to 0.92 for total test scores and 0.64 to 0.92 for subtest scores (5 of the 15 coefficients for subtests were below 0.70). Corresponding KR21 coefficients ranged from 0.89 to 0.92 for total test scores and 0.59 to 0.92 for subtest scores (6 of the 15 coefficients for subtests were below 0.70).

(2) Test-retest reliability: No information available.
(3) Alternate form reliability: Across test levels, alternate form reliability coefficients ranged from 0.62 to 0.88 for total test and subtest scores (2 of the 12 coefficients fell below 0.70).
(4) Inter-rater reliability: No information available.

Validity Evidence:
The SDRT 4 is a complete revision of the previous version, the SDRT 3 (published in 1986). In developing the SDRT 4, the authors reviewed the research literature, teacher surveys, curricula and instructional practices related to reading education, and difficulties identified in students with a reading diagnosis. They also reviewed the content, structure, and format of the SDRT 3 and then wrote all new items for the SDRT 4. The developers encourage test users to evaluate the alignment of the SDRT 4 to their school's instructional objectives and sequence. To support the developmental appropriateness of the format and content of the SDRT 4 Pink and Teal test levels, the authors point to increases in mean total test and subtest scores across grade-level increments and time (from fall to spring test administrations). They also present growth curves showing grade-level increases in median-scaled scores for subtests and the total test.

Construct/Concurrent validity: The authors report correlations for total test scores and subtest scores between the SDRT 4 and SDRT 3, respectively. For the Red through Blue forms, correlations between the two versions' total test scores ranged from 0.80 to 0.92. Across forms, correlations between subtests ranged from 0.76 to 0.82 (Phonetic Analysis); 0.58 to 0.77 (Vocabulary); 0.71 to 0.85 (Comprehension); and 0.43 to 0.57 (Scanning/Reading Rate).

The authors also present correlations between total test and subtest scores at adjacent test levels as evidence of continuity and consistency of SDRT 4 scores across levels. Pink and Teal level total scores correlated at 0.80, and correlations between corresponding subtest scores ranged from 0.42 to 0.68. The correlation between Teal and Red level total scores was 0.75, and correlations between subtests ranged from 0.52 to 0.58. For the Red through Blue levels, correlations between total scores across adjacent levels ranged from 0.80 to 0.87, and correlations between corresponding subtests on adjacent test levels ranged from 0.59 to 0.81. In addition, the authors report correlations between SDRT 4 Pink and Teal level total test and subtest scores with scores on the Stanford 10 Sounds and Letters subtest. These correlations ranged from 0.16 to 0.74.

The manual presents correlations between SDRT 4 scores and scores on the Otis-Lennon School Ability Test, Eighth Edition (OLSAT 8). Correlations between total test scores on the OLSAT 8 and the SDRT 4 Pink and Teal levels (fall and spring samples) ranged from 0.70 to 0.75; between SDRT 4 subtests and the OLSAT 8 Verbal subtest, correlations ranged from 0.26 to 0.67. Correlations between total test scores on the OLSAT 8 and the SDRT 4 Red through Blue forms ranged from 0.65 to 0.80 (only 1 of the 9 coefficients was below 0.70). Correlations between SDRT 4 subtests and the OLSAT 8 Verbal subtest ranged from 0.49 to 0.78 (18 of the 27 coefficients fell below 0.70).

At the Pink and Teal levels, correlations between the SDRT 4 Words, Stories, Sounds, and Letters subtests with the OLSAT 8 Nonverbal subtest ranged from 0.19 to 0.57. For the SDRT 4 Red through Blue levels as well, the same SDRT 4 subtests correlated to a lower degree with the OLSAT 8 Nonverbal subtest than with the OLSAT 8 Verbal subtest.

Predictive validity: No information available.

Bias Analysis: The developers of the SRDT 4 implemented two procedures to attempt to eliminate bias between gender, race, or ethnic groups. First, a panel of minority-group educators reviewed all SDRT 4 items for ethnic, gender, socioeconomic, cultural, and/or regional bias or stereotyping. Items identified as objectionable were deleted from the test or modified. Second, the developers used statistical procedures to examine differential item functioning between reference (males and White) and focal (female, Black, and Hispanic) groups of students matched on test scores. Items with a chi-square that exceeded what would normally be expected by chance were flagged for further review and possible elimination from the final test forms.

Training Support: Test kits for each level include booklets with detailed directions for administering each component of the test. For the Red through Blue levels, booklets detailing the administration directions are available for the Multiple Choice, Reading Questionnaire, Reading Strategies Survey, and Story Retelling portions of the test. Special training is not required, but assessors must thoroughly familiarize themselves with the test materials, procedures, and instructions before giving the test.

Adaptations/Special Instructions for Individuals with Disabilities: Large-print and Braille editions are available for visually-impaired students.

Alternate Forms: There are two alternate and equivalent forms, J and K, for the Purple, Brown, and Blue levels. Only one test form is available for the Pink through Green levels.

Previous Version: The SDRT 4 is a complete revision of the SDRT 3. The SDRT 3 consisted of four levels from the end of grade 1 through junior college. The fourth edition consists of eight levels from the beginning of kindergarten to the first semester of college. All SDRT 4 items were newly written during development of the measure.

NCEE or REL Study Use:[2] An Experimental Study of the Project Creating Independence through Student-Owned Strategies (CRISS) Reading Program (On 9th Grade Reading Achievement in Small Rural High Schools) (REL-Northwest)

[1] Individual ratings for internal consistency reliability coefficients for scores on subtests, clusters, and subclusters encompassed some ratings below the 0.70 level.

[2] See Table F.1 for web address.

References:

Engelhard, George. "Review of the Stanford Diagnostic Reading Test, Fourth Edition." In *The Thirteenth Mental Measurements Yearbook,* edited by James C. Impara and Barbara S. Plake. Lincoln, NE: Buros Institute of Mental Measurements, 1998.

Karlsen, Bjorn, and Eric Gardner. *Stanford Diagnostic Reading Test, 4th Edition*. *2004 Pink and Teal Norms Book and Technical Information*. San Antonio, TX: Harcourt Assessment, Inc., 2005.

Karlsen, Bjorn, and Eric Gardner. *Stanford Diagnostic Reading Test, 4th Edition*. *1995 Multilevel Norms Book and Technical Information*. San Antonio, TX: Harcourt, Inc., 1996.

Karlsen, Bjorn, and Eric Gardner. *Stanford Diagnostic Reading Test Screening Test, 4th Edition*. *2004 Pink and Teal Norms Book and Technical Information*. San Antonio, TX: Harcourt Assessment, Inc., 2005.

TERRANOVA 3, 2008

Authors: CTB/McGraw-Hill		**Type of Assessment:** Group-administered assessment **Domains:** Reading, language arts/language proficiency, mathematics, science, and social studies
Publisher: CTB/McGraw-Hill 800-538-9547 http://www.ctb.com		**Grade/Age Range:** Kindergarten through grade 12 **Administration Interval:** Annual
Materials, Training, and Scoring Costs: *Complete Battery*: Test per 25 booklets: $190 (levels 10 through 13), $140 (levels 14 through 22); *Plus Supplement*: $63.25 (levels 10 through13), $54.50 (levels 14 through 22); 50 answer sheets: $63 (levels 14 through 22) *Survey*: Test per 25 booklets: $148 (levels 12 and 13), $133.25 (levels14 through 22); *Plus Supplement*: $63.25 (levels 11 through 22); 50 answer sheets: $63 (levels 14 through 22) *Multiple Assessments*: Test per 25 booklets: $198.50; *Plus Supplement*: $63.25 (levels 11 through 13), $54.50 (levels 14 through 22); Manipulatives: $17.35 per 25 (levels 13 through 22) Scoring guide: $132.50 Teacher's guide: $67.20 Test directions for teachers are included with test booklet orders for each level.		**Personnel and Training Requirements** **Credentials Required for Use:** Level 1 (training or supervised experience with measurement) **Personnel for Administration:** Individual with basic clerical skills and some training **Training for Administration:** Self-training (<1 hour)
Languages: English and Spanish		**Alternate Forms:** No
Representativeness of Norming Sample: The norming sample[11] was gathered in 2006 and 2007 from a nationally representative sample of about 200,000 students in kindergarten through grade 12. Stratification was based on school type (public, private, or parochial), geographic region, community type, and socioeconomic status. The sample included students with disabilities.		**Summary** **Initial Cost:** 3 ($200 to $500) **Time to Administer:** Approximately 2 to 5 hours depending on battery module (see Description) **Ease of Administration and Scoring:** 4 (administered or scored by a clinician or specialist) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 3 (normed in past 10 years and nationally representative)

NARRATIVE

Description: The TerraNova 3 is a group-administered assessment of reading, language, mathematics, science, and social studies. These areas are each assessed with subtests through three types of modules (Complete Battery, Survey, and Multiple Assessments). An optional Plus module, available in the TerraNova, The Second Edition may be used in combination with any of the three modules. Subtests in the Plus module include Word Analysis, Spelling, Vocabulary, Language Mechanics, and Mathematics Computation. Modules include several testing levels appropriate for kindergarten through grade 12, and testing levels overlap between grades to permit a vertical scale for reporting and comparing results across grade levels, as shown below. Students complete the test in a specified amount of time, which varies according to module and grade level. Students fill in their answers in test booklets with pencil and use calculators if they are accustomed to using them in the classroom. Mathematics manipulatives (rulers) are available for testing levels 13 through 22.

Level	Grade
10	K.6 to 1.6
11	1.6 to 2.6
12	2.0 to 3.2
13	2.6 to 4.2
14	3.6 to 5.2
15	4.6 to 6.2
16	5.6 to 7.2
17	6.6 to 8.2
18	7.6 to 9.2
19	8.6 to 10.2
20	9.6 to 11.2
21–22	10.6 to 12.9

Note: The decimal number after the grade is the number of months that have elapsed in the school year.

Each module of the TerraNova 3 (Complete Battery, Survey, and Multiple Assessment) varies in terms of subtests, number of items, and average administration time (Tables 1 through 3). The Complete Battery assesses students in kindergarten through grade 12 (using levels 10 through 22) and includes different subtests according to testing level, with the number of items ranging from 70 to 217. Subtests include Reading and Mathematics (level 10); Reading, Mathematics, Science, and Social Studies (levels 11 and 12); and Reading, Language, Mathematics, Science, and Social Studies (levels 13 through 22). The Complete Battery requires 1.5 to over 4 hours of child test-taking time (Table 1), excluding 20 to 45 minutes of additional administration time. The Plus subtests significantly increase the number of items and lengthen the testing time by 50 minutes to up to 1 hour, 20 minutes.

The Survey is shorter than the Complete Battery but still permits criterion-referenced interpretation of scores and performance-level data for grades 2 through 12 (levels 12 through 22). Subtests include Reading, Language, Mathematics, Science, and Social Studies in levels 13 through 22. The level 12 assessment excludes Language. The Survey includes between 106 and 142 items and takes from 2 hours, 15 minutes to 2 hours, 50 minutes for completion (Table 2).

The Plus subtests significantly increase the number of items and lengthen the testing time from 1 hour, 5 minutes to up to 1 hour, 20 minutes.

The Multiple Assessments are administered to students in grades 1 through 12 (levels 11 through 22). Subtests include Reading, Language, Mathematics, Science, and Social Studies for levels 12 through 22; level 11 and 12 assessments exclude Language. These tests are much longer, containing 136 to 184 items, and require between 4 hours, 20 minutes and 5 hours, 35 minutes of child test-taking time (Table 3), excluding 20 to 45 minutes of additional administration time. The Plus subtests significantly increase the number of items and lengthen testing time by 50 minutes to up to 1 hour, 5 minutes.

Table 1. Complete Battery and Plus

Testing Time in Hours: Minutes

Level	Complete Battery	Plus	Reading	Language	Mathematics	Science	Social Studies	Number of Items (Complete Battery)
10	1:35	n.a.	0:55	n.a.	0:40	n.a.	n.a.	70
11	2:40	0:50	0:55	n.a.	1:05	0:20	0:20	127
12	3:00	1:20	1:10	n.a.	1:00	0:25	0:25	149
13	4:05	1:20	1:00	0:35	1:10	0:40	0:40	190
14– 15	4:05	1:05	1:00	0:35	1:10	0:40	0:40	217
16	4:05	1:05	1:00	0:35	1:10	0:40	0:40	216
17	4:05	1:05	1:00	0:35	1:10	0:40	0:40	217
18	4:05	1:05	1:00	0:35	1:10	0:40	0:40	216
19– 20	4:05	1:05	1:00	0:35	1:10	0:40	0:40	206
21–22	4:05	1:05	1:00	0:35	1:10	0:40	0:40	206

Source: Publisher web site at http://www.ctb.com.
n.a. = not applicable.

Table 2. Survey and Plus

Testing Time in Hours: Minutes

Level	Survey	Plus	Reading	Language	Mathematics	Science	Social Studies	Number of Items (Survey)
12	2:15	1:20	1:00	n.a.	0:35	0:20	0:25	106
13	2:50	1:20	0:50	0:30	0:40	0:25	0:25	125
14–15	2:50	1:05	0:50	0:30	0:40	0:25	0:25	142
16	2:50	1:05	0:50	0:30	0:40	0:25	0:25	141
17	2:50	1:05	0:50	0:30	0:40	0:25	0:25	142
18	2:50	1:05	0:50	0:30	0:40	0:25	0:25	141
19–22	2:50	1:05	0:50	0:30	0:40	0:25	0:25	135

Source: Publisher web site at http://www.ctb.com.
n.a. = not applicable.
Note: The survey is not available for use with students in levels 10 and 11.

Table 3. Multiple Assessments and Plus

Testing Time in Hours: Minutes

Level	Multiple Assessments	Plus	Reading	Language	Mathematics	Science	Social Studies	Number of Items (Multiple Assessments)
11	4:30	0:50	1:45	n.a.	1:15	0:45	0:45	136
12	4:20	1:30	1:35	n.a.	1:15	0:45	0:45	138
13	5:35	1:20	1:20	0:40	1:30	1:00	1:05	153
14	5:35	1:05	1:20	0:40	1:30	1:00	1:05	183
15	5:35	1:05	1:20	0:40	1:30	1:00	1:05	182
16	5:35	1:05	1:20	0:40	1:30	1:00	1:05	184
17	5:35	1:05	1:20	0:40	1:30	1:00	1:05	181
18	5:35	1:05	1:20	0:40	1:30	1:00	1:05	180
19	5:35	1:05	1:20	0:40	1:30	1:00	1:05	175
20	5:35	1:05	1:20	0:40	1:30	1:00	1:05	176
21–22	5:35	1:05	1:20	0:40	1:30	1:00	1:05	174

Source: Publisher web site at http://www.ctb.com.
n.a. = not applicable.
Note: The Multiple Assessments are not available for use with students in level 10.

Other Languages: The Supera contains Spanish-language versions of three modules in the TerraNova, The Second Edition and is available for students in grades 1 through 10. The Supera Survey and Multiple Assessments modules assess Reading, Language, and Mathematics. The Supera Plus module, which may be added to the survey or the Multiple Assessments, assesses Word Analysis, Vocabulary, Language Mechanics, Spelling, and Mathematics. The Supera Multiple Assessments and Survey modules were normed with Spanish-speaking U.S. students in 1999 and 2000. Equivalence data between the Supera and the TerraNova, The Second Edition are not available.

Uses of Information: The TerraNova 3 was designed to assess individual student achievement in several domains and to track student progress in relation to a nationally representative sample. The publisher states that the TerraNova 3 results may be used as a diagnostic assessment to predict outcomes on state assessments required by No Child Left Behind and to identify at-risk students and schools in order to target areas for improvement. In a few states, the TerraNova 3 is customized to meet state testing restrictions, such that particular levels are not available for testing. The publisher maintains a continuously updated list of these states and notifies purchasers of testing restrictions as applicable.

Methods of Scoring: Selected-response items (i.e., students choose a response) are electronically scanned while the publisher's professional staff follows rubrics to score constructed-response items (i.e., students provide short or extended responses). The publisher reports two types of scale scores; users may select either the number of correct responses (i.e., raw scores) or scores based on Item Response Theory (IRT) scaling. The latter scores factor in the psychometric characteristics of each item, such as level of difficulty, with estimates based on the entire pool of items for a given content area even though students answer only a subset of the items. Developers have noted for previous TerraNova versions that IRT scores are more reliable and less susceptible to error than raw scores (CTB/McGraw-Hill 2003). Scale scores provide the

basis for the inclusion in norm-referenced scores of percentiles, stanines, grade equivalents, and normal curve equivalents. Criterion-referenced scores are available with the Objective Performance Index (OPI) scores (i.e., low, moderate, and high degrees of mastery) for knowledge or skill areas within subtests (e.g., Analyzing Text within the Reading subtest).

Interpretability: Researchers may use scale scores for statistical analysis (CTB/McGraw-Hill 2003). The publisher provides normative and criterion performance reports of individual students and groups. Individual student reports target areas of instructional need and are available for school staff and families. An additional "translation guide" assists families in interpreting their children's scores compared with other students. Guides are available in 10 languages to assist non-English-speaking families. Several types of group reports are available for teachers to help them track progress in classes and for administrators to help them track progress across grade levels and within the school. Student-level reports are available at an additional cost per student.

Reliability:
(1) Internal consistency reliability: Cronbach's alpha coefficients for subtest and composite scores of the TerraNova 3 were calculated in 2006 and 2007 for Reading, Language, and Mathematics; by test level for the Complete Battery, Survey, and Multiple Assessment modules. Coefficients for scores from the Complete Battery, Survey, and Multiple Assessments modules ranged from 0.72 to 0.97, from 0.74 to 0.95, and from 0.81 to 0.96, respectively. Reliability estimates for scores from most individual subtests were in the 0.80s and 0.90s. Reliability estimates for composite scores for all levels were in the 0.90s.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: Developers calculated intraclass and weighted Kappa correlation coefficients in 1999 and 2000 for the TerraNova, The Second Edition for 434 constructed-response items of Multiple Assessment Form C by subtest (Reading and Language, Mathematics, Science, and Social Studies) and grade (1 through 12). Modules for each grade included 33 to 43 constructed-response items with between 405 and 866 student responses. Ranges in the number of items and students reflect variation by subtest and grade. Intraclass correlation coefficients, which did not account for chance agreement between raters, ranged from 0.78 to 1.0. Weighted Kappa coefficients, which took chance agreement into account, ranged from 0.70 to 1.0 for 425 of 434 items. Nine of 434 items had coefficients ranging from 0.57 to 0.67.

Validity Evidence:
Planning for assessment content involved meetings of advisory panels, including teachers, administrators, and content specialists. Developers conducted reviews of curricula and content standards across several states and districts and reviewed textbook series, practices of model education programs, and publications of national academic standards. Professionals, mainly teachers, developed a pool of potential items and collaborated with artists and graphic designers to ensure graphic and textual clarity. In a departure from development of the TerraNova, The Second Edition, content developers increased the rigor of Reading, Language, and Mathematics items and maximized item alignment with state curricula. In particular, developers modeled item development and reading passages for the Reading and Language subtests on the National Assessment of Educational Progress (NAEP). In addition, they designed the Mathematics subtest to be similar to NAEP by emphasizing problem solving, communication, reasoning, and connections. During the tryout testing phase in 2005, developers collected information from

nearly 44,000 students by grade level on test clarity and appropriateness of content. The tryout included items from the TerraNova, The Second Edition as well as corresponding new items by subtest and level. To obtain nationally representative estimates, developers linked new items to the scales based on the norming sample calibrations. Developers selected items based on the items' psychometric characteristics estimated by IRT scaling.

Construct/Concurrent Validity: Developers correlated scores from the TerraNova 3 with scores from the InView,[1] an assessment of academic ability with five subtests: Verbal Reasoning-Words, Verbal Reasoning-Context, Sequences, Analogies, and Quantitative Reasoning. Correlations were estimated for total scores and each subtest by level.

Correlations between InView total verbal scores and TerraNova 3 composite scores of Reading and Vocabulary ranged from 0.61 to 0.77. Correlations between InView total nonverbal scores and TerraNova 3 composite Mathematics scores ranged from 0.54 to 0.77.

Correlations between InView total verbal scores and TerraNova 3 composite Mathematics scores ranged from 0.52 to 0.73 while the coefficients between InView nonverbal scores and TerraNova 3 composite Reading scores and composite Language scores ranged from 0.52 to 0.68 and from 0.47 to 0.68, respectively.

Within the TerraNova 3, the composite Mathematics score demonstrated correlations of 0.66 with the composite Reading Score, 0.60 with the Language score, and 0.54 with the Spelling score for grade 2 students. For students in grades 3 through 12, coefficients ranged from 0.55 to 0.72, from 0.54 to 0.78, and from 0.38 to 0.64, respectively. Similarly, the Science score was correlated with the composite Reading, Language, and Spelling scores. For grade 2 students, correlations were 0.48, 0.37, and 0.21, respectively. For students in grades 3 through 12, correlations ranged from 0.48 to 0.77, from 0.37 to 0.73, and from 0.38 to 0.49, respectively.

For two subtests, Science and Social Studies, developers correlated scores for forms between the TerraNova 3 and the TerraNova, The Second Edition. Correlations for levels 11 through 20 (data unavailable for levels 21 and 22) ranged from 0.51 to 0.78 for Science and from 0.61 to 0.78 for Social Studies. Developers noted that the two forms for each subtest were administered at the same time, with no further elaboration.

Predictive validity: No information available.

Bias Analysis: Developers conducted differential item functioning (DIF) analyses with nearly 44,000 students during the tryout testing phase in 2005. They compared two subgroups, including gender and ethnicity (Hispanic, Black, and Other, which included non-Hispanic and non-Black students) and examined the items for the TerraNova 3 and previous TerraNova editions. They did not, however, report whether any items showed bias or whether items were removed. Developers also avoided item bias by considering item comments from educators across the country with different perspectives regarding language, subject matter, and group representation.

Training Support: The Teacher's Guide provides clear and easy-to-understand directions for administration.

Adaptations/Special Instructions for Individuals with Disabilities: Materials are available in Braille, although the content of the Braille version is not identical to the regular assessment because some items do not lend themselves to translation into Braille. Braille adaptations have not been normed. Although supporting documentation for the TerraNova 3 does not describe other adaptations, the TerraNova, The Second Edition includes adaptations such as modification of the presentation (e.g., large print), response process (e.g., responses to a scribe), test setting (e.g., alone in a study carrel), timing and scheduling (e.g., more breaks), and administration (e.g., oral). The publisher provides guidance on how to interpret test scores when the TerraNova, The Second Edition is administered with varying degrees of accommodations (CTB/McGraw-Hill 2005).

Alternate Forms: None.

Previous Version: The TerraNova 3 updates the first TerraNova (also called the CTBS), published in 1997, and the TerraNova, The Second Edition (also called the California Achievement Tests 6th Edition, or CAT/6), published in 2001. The TerraNova 3 includes new items and updated norms but measures the same constructs in the same manner as the first and second editions. The main difference between the TerraNova 3 and the second edition relates to the Reading and Language Arts subtests. The TerraNova 3 separates the Reading and Language subtests into two for levels 13 through 22. In addition, it assesses phonics and phonemic awareness in the Reading subtest rather than in Language for kindergarten through grade 2. For Reading, Language, and Mathematics, content developers increased item rigor and alignment with state curricula.

NCEE or REL Study Use:[2] Effects of Odyssey Math® Software on the Mathematics Achievement of Selected Fourth Grade Students in the Mid-Atlantic Region: A Multi-Site Cluster Randomized Trial (REL-Mid-Atlantic); The Effect of Connected Mathematics Program 2 (CMP2) (on the Math Achievement of Middle School Students in Selected Schools in the Mid-Atlantic Region) (REL-Mid-Atlantic); Efficacy of Frequent Formative Assessment for Improving Instructional Practice and Student Performance, Given Variations in Training to Use Assessment Results (REL-Midwest); Intensive Small Group Math Study (REL-Southwest); An Evaluation of Teachers Trained through Different Routes to Certification.[3]

[1] The TerraNova 3 was co-normed with the InView (grades 2 through 12) and the Primary Test of Cognitive Skills (kindergarten and grade 1) in 2006 and 2007 (CTB/McGraw-Hill 2008).

[2] See Table F.1 for web address.

[3] This study used the California Achievement Test, Fifth Edition (CAT/5), which was a previous version of the CAT/6 (also called the TerraNova, The Second Edition). The TerraNova 3 has replaced the CAT series such that there will be no future CAT editions.

References:

Cizek, Gregory J. "Review of the TerraNova, the Second Edition." In *The Sixteenth Mental Measurements Yearbook*, edited by Robert Spies and Barbara Plake. Lincoln, NE: Buros Institute of Mental Measurements, 2005.

CTB/McGraw-Hill. "Beyond the Numbers: A Guide to Interpreting and Using the Results of Standardized Achievement Tests." Monterey, CA: CTB/McGraw-Hill Companies, Inc., 2003.

CTB/McGraw-Hill. "Guidelines for Inclusive Test Administration." Monterey, CA: CTB/McGraw-Hill Companies, Inc., 2005.

CTB/McGraw-Hill. "Inform: Why Test in the Fall?" CTB/McGraw-Hill Companies, Inc., 2004.

CTB/McGraw-Hill. *SUPERA*. Monterey, CA: CTB/McGraw-Hill Companies, Inc., 1997.

CTB/McGraw-Hill. *Teacher's Guide to TerraNova, Third Edition*. Monterey, CA: CTB/McGraw-Hill Companies, Inc., 2008.

CTB/McGraw-Hill. "TerraNova, the Second Edition: Frequently Asked Questions." Monterey, CA: CTB/McGraw-Hill Companies, Inc., 2000.

CTB/McGraw-Hill. "TerraNova, Third Edition: Technical Bulletin 1." Monterey, CA: CTB/McGraw-Hill Companies, Inc., 2008.

TEST OF EARLY MATHEMATICS ABILITY, THIRD EDITION (TEMA-3), 2003

Authors: Herbert P. Ginsburg and Arthur J. Baroody		**Type of Assessment:** Individually administered adaptive assessment **Domain:** Mathematics
Publisher: PRO-ED, Inc. 800-897-3202 http://www.proedinc.com		**Grade/Age Range:** 3 years through 8 years, 11 months **Administration Interval:** None described
Material, Training, and Scoring Costs: TEMA-3 Kit (includes Examiner's Manual, Picture Book Form A, Picture Book Form B, Examiner Record Booklets Form A (25), Examiner Record Booklets Form B (25), Worksheets Form A (25), Worksheets Form B (25), Assessment Probes and Instructional Activities, and manipulatives): $278		**Personnel and Training Requirements** **Credentials Required for Use:** Level 1 (training or supervised experience with measurement) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (> 2 hours) Although formal training in psychometrics is not required, the assessor should be thoroughly familiar with the test materials and well trained in administering and scoring tests. In addition, the assessor is advised to practice administering the test to at least 3 students.
Languages: English		**Alternate Forms:** Two forms; administration interval not described.
Representativeness of Norming Sample: The norming sample consisted of a sample of 1,228 students age 3 through 8 years (about 100 each of 3- and 4-year-olds and about 200 each of 5- through 8-year-olds) attending child care centers or general education classes, based on publisher customer records. The characteristics of the sample were compared to the 1999 U.S. Census Bureau's *Statistical Abstract of the United States*, and the sample was then weighted at each age level based on region, ethnicity, and gender in order to be more proportional to the national school-age population. The sample includes students with disabilities if enrolled in general education classes. Norming was conducted between fall 2000 and spring 2001 in 15 states representing the four major U.S. regions.		**Summary** **Initial Material Cost:** 3 ($200 to $500) **Time to Administer:** 45 to 60 minutes **Ease of Administration and Scoring:** 3 (administered and scored by trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 3[1] (normed within past 10 years and nationally representative)

NARRATIVE

Description: The TEMA-3 is an individually administered adaptive assessment designed to assess the informal and formal mathematical concepts and skills of students (age 3 years through 8 years, 11 months). The assessment includes an easel administration with the use of manipulatives for some items. Each of two parallel forms, Forms A and B, includes 72 items, with each item consisting of one to six questions. The assessment is untimed, with a reported average testing time of 45 to 60 minutes. The items are arranged in order of increasing difficulty, and floor and ceiling rules are used to determine the items administered to each student. The assessor begins the assessment at the entry point established for each age group and administers the item sets until the student's "ceiling" is established (the student incorrectly answers five items in a row). The basal is then established by identifying the five consecutive correct responses closest to the ceiling.

The publisher does not discuss any floor or ceiling effects, but a review by Bliss (2006) notes that an examination of the raw score conversion table suggests floor effects with students under 4 years, 3 months of age and potential ceiling effects for the three highest age ranges (8 years, 3 months through 8 years, 11 months). For example, the reviewer notes that a 3-year-old student can earn no points and still receive a Math Ability Score of 85, 1 standard deviation below the mean.

Other Languages: None.

Uses of Information: The publisher states that the TEMA-3 may be used to screen for giftedness and developmental delays and to measure a student's mathematical strengths and weaknesses. Suggested followup probes and instructional activities are available. The TEMA-3 may also track student's progress in acquiring mathematical knowledge.

Methods of Scoring: For each item, the Profile/Examiner Record Booklet indicates how many questions the student must answer correctly to earn credit for the item. The assessor may compute a raw score and, using tables provided in the manual, covert the raw score into an age-equivalent, grade-equivalent, percentile rank, and standard score referred to as a Math Ability Score. An individual familiar with the TEMA-3 scoring criteria and well trained in test administration and scoring should score the assessment. The raw score is the total of all correctly answered items. All items below the basal are scored as correct and those above the ceiling as incorrect. The Math Ability Scores at three-month intervals were estimated by using polynomial regression and "smoothed somewhat to allow for a consistent progression across age levels" (Ginsburg and Baroody 2003). Given the restricted range of the standardized score, the publisher recommends use of the assessment as a criterion-referenced measure for 3-year-old students.

Interpretability: The publisher recommends assessor training in test administration. The manual includes a chapter addressing interpretation issues.

Reliability:
(1) Internal consistency reliability: Based on data from the entire norming sample, Cronbach's alphas ranged from 0.92 to 0.95 for Form A and from 0.95 to 0.96 for Form B across the six age intervals from 3- through 8-years-old.

(2) Test-retest reliability: Form A was administered to 49 students 3- through 8-years-old from New York and North Dakota. The correlation between scores on two administrations (with a two-week interval between sessions) was 0.82. Form B was administered to 21 students 4- through 8-years-old from North Dakota. The correlation between scores on two administrations was 0.93. The correlations for both Form A and Form B were corrected for restricted range.
(3) Alternate form reliability: Two alternate form reliability analyses were conducted with 46 5- through 8-year-old students from Texas. First, each student was administered two forms of the assessment in a counterbalanced design during the same testing session. The scores correlated 0.97. Second, two weeks later, both forms were re-administered to each student in the opposite order from the previous administration. The correlation coefficient was 0.93. Both coefficients were corrected for restricted range.
(4) Inter-rater reliability: No information available.

Validity Evidence:
The authors discuss in detail the rationale and research base for the items selected for the assessment. They describe the development of children's informal mathematics skills in numbering, number conception, calculation, and the understanding of concepts such as the cardinality rule. They also review children's formal mathematics development including numeral literacy, mastery of number facts, calculation skills, and understanding of concepts such as additive commutativity.

Construct/Concurrent validity: The authors conducted item discrimination and item difficulty analyses with data from the entire norming sample. The item discrimination indices ranged from 0.45 to 0.66 for Form A and from 0.53 to 0.68 for Form B across the six age intervals from 3- through 8-years-old. The median item difficulties ranged from 0.04 to 0.67 for Form A and from 0.03 to 0.87 for Form B across age intervals.

Scores on the TEMA-3 were compared to the mathematical ability portions of the following assessments: KeyMath-Revised/Normative Update, Woodcock-Johnson III Tests of Achievement, Diagnostic Achievements Battery-Third Edition, and Young Children's Achievement Test. The correlation coefficients ranged from 0.54 to 0.91 after correcting for the restricted range and measurement error (due to lower reliability in some of the criterion measures). The sample sizes for each assessment ranged from 43 to 62 students.

Additionally, the mean scores of the assessment increased with the age of the student, and the scores for students identified as low mathematics achievers were below average.

Predictive validity: No information available.

Bias Analysis: For seven subgroups (male, female, White, Black, Hispanic, Asian, and students with low mathematics achievement), internal consistency ranged from 0.98 to 0.99 across forms and groups. The standard scores for selected subgroups (male, female, White, Black, and Hispanic) fell within the average range. Differential item functioning was assessed by comparing the scores of three groups (male versus female, Black versus non-Black, and Hispanic versus non-Hispanic) to those of the entire normative sample across all items for each form. Five item comparisons were statistically significant at the $p < 0.001$ level. On Form A, two significant indices of bias were identified for the Black/non-Black group and one for the male/female group.

For Form B, two indexes of bias were identified as significant for the male/female group. The effect sizes were negligible, and thus items were not removed.

Training Support: No information available.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: The TEMA-3 has two parallel forms—A and B. A linear equating procedure was used to correct for any differences in difficulty between the two forms. Testing intervals between forms were not described.

Previous Version: The TEMA-3 updates the TEMA-2 published in 1990. Additional items were added to provide a more comprehensive assessment. An alternate form was introduced with demonstrated equivalence. Directions for administering and scoring the items were clarified. Manipulatives necessary for administration were added; Picture Books were printed in an easel-back format and in color; the names of some items were clarified; and the trial numbering system was revised.

NCEE or REL Study Use:[2] Intensive Small Group Math Instruction Study (REL-Southwest)

[1] The sample was weighted to be proportional to the national school-age population.

[2] See Table F.1 for web address).

References:

Bliss, Stacy. "Review of Test of Early Mathematics Ability-Third Edition." *Journal of Psychoeducational Assessment*, vol. 24, no. 1, 2006, pp. 85-88.

Crehan, Kevin D. "Review of the Test of Early Mathematics Ability, Third Edition." In *The Sixteenth Mental Measurements Yearbook*, edited by Robert A. Spies and Barbara S. Plake. Lincoln, NE: The Buros Institute of Mental Measurements, 2005.

Ginsburg, Herbert, and Arthur J. Baroody. *TEMA 3: Test of Early Mathematics Ability--Third Edition Examiner's Manual*. Austin, TX: PRO-ED, Inc., 2003.

Monsaas, Judith A. "Review of the Test of Early Mathematics Ability, Third Edition." In *The Sixteenth Mental Measurements Yearbook*, edited by Robert A. Spies, Barbara S. Plake, and Linda L. Murphy. Lincoln, NE: The Buros Institute of Mental Measurements, 2005.

TEST OF ECONOMIC LITERACY, THIRD EDITION (TEL-3), 2001

Authors: William B. Walstad and Ken Rebeck		**Type of Assessment:** Group-administered assessment **Domain:** Social studies
Publisher: National Council on Economic Education 800-338-1192 http://www.ncee.net		**Grade/Age Range:** Grades 9 through 12 **Administration Interval:** As frequently as 3 times a semester
Material, Training, and Scoring Costs: Set of 25 test booklets (Form A or Form B): $22.95 Examiner's Manual: $17.95		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2 (bachelor's degree with coursework in measurement and domain) **Personnel for Administration:** Individual with basic clerical skills with some training **Training for Administration:** Self-training (<1 hour)
Languages: English		**Alternate Forms:** Two forms; administered as frequently as 3 times a semester
Representativeness of Norming Sample: The norming sample was a sample consisting of 7,243 students (459 10th graders, 1,789 11th graders, and 4,213 12th graders)[1] from 100 high schools nationwide. Approximately 81 percent had had economics instruction at the time of testing. The tests were conducted at the end of fall semester 1999 and the end of spring semester 2000.		**Summary** **Initial Material Cost:** 1 (under $100) **Time to Administer:** 30 to 40 minutes **Ease of Administration and Scoring:** 2 (self-administered or administered and scored by someone with basic clerical skills) **Reliability:** 3[2] (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available[2] **Norming Sample Characteristics:** 2 (older than 10 years or not nationally representative)

NARRATIVE

Description: The Test of Economic Literacy, Third Edition, is a group-administered achievement assessment designed to measure understanding of basic economic concepts (fundamental economic concepts, microeconomic concepts, macroeconomic concepts, and international economic concepts) for students in grades 9 through 12. Students receive a test booklet and answer form and have 40 minutes to complete the paper-and-pencil assessment consisting of 40 multiple-choice questions.

Other Languages: None.

Uses of Information: The TEL-3 is an achievement assessment to measure high school students' understanding of basic economic concepts (fundamental economic concepts, microeconomic concepts, macroeconomic concepts, and international economic concepts). The authors also note that it may be used as a pre-test to adjust curriculum, as a post-test to measure improvement and differences by student subgroup, as a mid-course evaluation to aid instruction, as a pre-test for college-level instruction, and as a research tool.

Methods of Scoring: The raw score is the total number of questions answered correctly. It may be converted into a percentile rank by using tables in the examiner's manual based on form, class type, and economics experience. Some percentile rankings, however, are based on small samples (fewer than 300 students). Performance on individual items may be compared to item difficulty in the norming sample. In addition, for those students who have had economics instruction, a table equates the raw scores of the alternate forms.

Interpretability: The assessment is designed for administration and interpretation by high school teachers or administrators. Several tables in the manual are intended to help assessors analyze students' responses to items and alter curriculum based on students' perceived understanding.

Reliability:[2]
(1) Internal consistency reliability: The Cronbach's alpha coefficient for both Forms A and B was 0.89.
(2) Test-retest reliability: The correlation coefficient between students' scores on the first and second administrations (interval described as "a short time period") of Form A (N = 37) was 0.94. Test-retest reliability was not evaluated for Form B.
(3) Alternate form reliability: No information available.
(4) Inter-rater reliability: No information available.

Validity Evidence:[2]
Development of items for the TEL-3 was based on the *Framework for Teaching Basic Economic Concepts* (1995) and the *Voluntary National Content Standards in Economics* (1997) and grouped into 21 economic concepts outlined in the *Framework*. The 21 concepts fall into four broad categories, and a certain percentage of items fall into each of the four categories: fundamental economic concepts (35 percent), microeconomic concepts (25 percent), macroeconomic concepts (25 percent), and international economic concepts (15 percent). Items also were developed to vary across three levels of cognition: knowledge, comprehension, and

application. The TEL-3 draft was sent to three national committees for professional judgment/input on content, areas of potential bias, and potential for reading problems. The committees comprised experienced high school economics teachers, economists and educators who serve as directors of economic centers or councils, and distinguished economists working to improve economic education. Committee feedback was incorporated into the measure, and field testing was then conducted for item difficulty and testing administration problems.

Construct/Concurrent validity: The authors present the feedback and field test results as evidence of the TEL-3's content validity for measuring general achievement in basic economic concepts, not as a test of mastery of any or all of the 21 concepts. In addition, the authors note that analysis of the raw scores from the norming sample indicates a significant difference between students with and without economics instruction. However, across the norming sample, 81 percent of students had economics instruction, largely reflecting the over-representation of 12th graders (94 percent of the grade 12 students had economics instruction).[1]

Authors correlated the TEL-3 with the Test of Economic Literacy, Second Edition (TEL-2) scores. The correlation between scores on TEL-2 and Form B TEL-3 was 0.81 for students in regular economics classes (N = 23). The correlation between scores on TEL-2 and Form A TEL-3 was 0.67 only for those students in Advanced Placement (AP) or honors economics classes (N = 16). The authors also examined the results of the TEL-3 for those students (n = 68) who scored a 3 or above on the AP economics examinations. Students who scored a 5 on the AP examinations scored on average two points higher (39 out of 40) on the TEL-3 than students who scored a 3 on the AP examinations (37 out of 40). No statistical test results were provided.

The authors grouped a subsample of 4,613 students according to verbal ability (low, middle, high) by using an adapted vocabulary test to produce groups of "sufficient" size to make "rough" comparisons across ability levels. The authors examined the relationship with economics instruction after controlling for general verbal ability (vocabulary). The authors note that, despite differences in verbal ability, exposure to economics instruction also made a "significant" difference within each verbal ability level. No statistical test results were provided.

Authors tested for subgroup differences using regression analysis (controlling for verbal ability) with data from the norming sample. They found a significant difference on the performance on the TEL-3 between those students with and without economic instruction. Additional regression analyses found similar results when controlling for "other student-specific and demographic" data.

Predictive validity: No information available.

Bias Analysis: Professional committees reviewed items for bias in content and wording. In addition, the TEL-2 assessment underwent an analysis of differential item functioning (DIF). The manual notes this work as part of the development of the TEL-3 but provided no results. DIF analysis was not conducted for the TEL-3.

Training Support: No information available.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: The TEL-3 has two alternate forms—Form A and Form B. The authors describe a potential testing scenario in which students are given the TEL-3 at the beginning, middle, and end of a semester, noting that forms should be alternated for this or shorter administration periods.

Previous Version: The TEL-3 was updated to cover more content considered to be basic economic concepts.

NCEE or REL Study Use:[3] High School Instruction with Problem-Based Economics (REL-West)

[1] Information from Examiner's Manual Tables 21 and 22, pp. 30-31; no information provided as to why numbers across grades do not sum to sample total.

[2] Reliability and validity evidence was conducted with samples of students, not teachers, but the measure has been used with both groups in some studies.

[3] See Table F.1 for web address.

References:

Walstad, William B., and Ken Rebeck. *Test of Economic Literacy, Third Edition: Examiner's Manual*. New York: National Council of Economic Education, 2001.

Young, John W. "Review of the Test of Economic Literacy." In *The Fifteenth Mental Measurements Yearbook*, edited by Barbara S. Plake, James C. Impara, Robert A. Spies. Lincoln, NE: Buros Institute of Mental Measurements, 2003.

TEST OF LANGUAGE DEVELOPMENT-PRIMARY, FOURTH EDITION (TOLD-P:4), 2008

Authors: Phyllis L. Newcomer and Donald D. Hammill		**Type of Assessment:** Individually administered adaptive assessment **Domain:** Language arts/language proficiency (expressive and oral receptive language; syntax; semantics; phonology)
Publisher: PRO-ED, Inc. 800-897-3202 http://www.proedinc.com		**Grade/Age Range:** 4 years through 8 years, 11 months **Administration Interval:** Once or twice yearly
Material, Training, and Scoring Costs: TOLD-P:4 kit (includes a sturdy storage box containing the Examiner's Manual, Picture Book, 25 Examiner/Record Forms, and a *Critical Reviews and Research Findings for TOLD-P: 1977–2007* monograph): $299		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2 (bachelor's degree with coursework in measurement and domain) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (> 2 hours) Assessors must be formally trained in assessment and evaluation of language abilities.
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: The TOLD-P:4 was normed on a sample of 1,108 students from 16 states selected to represent the four major regions of the United States. The sample included 166 4-year-olds, 182 5-year-olds, 268 6-year-olds, 266 7-year-olds, and 226 8-year-olds. Demographic characteristics (gender, geographic region, race/ethnicity, exceptionality status, family income, and educational level of parents) were representative of the *Statistical Abstract of the United States* for the 2005 school-age population and were stratified by age (excluding exceptionality status). Testing took place during winter 2006 through fall 2007.		**Summary** **Initial Material Cost:** 3 ($200 to $500) **Time to Administer:** Approximately 35 to 50 minutes core subtests; supplemental subtests additional 30 minutes during a separate testing session. **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 3 (normed within past 10 years and nationally representative)

NARRATIVE

Description: The TOLD-P:4 is an individually administered adaptive assessment that measures spoken language of 4- to 8-year-olds. Administration requires an easel for particular subtests, as students must identify images orally or by pointing. Other subtests require students to respond to questions, provide definitions, repeat after the assessor, provide the missing word in a sentence, distinguish between words, or break words apart based on questions or prompts from the assessor (all performed orally). The assessment comprises nine subtests (six core and three supplemental) with a total of 285 items measuring various aspects of oral language. The six core subtests are (1) Picture Vocabulary (34 items), (2) Relational Vocabulary (34 items), (3) Oral Vocabulary (38 items), (4) Syntactic Understanding (30 items), (5) Sentence Imitation (36 items), and (6) Morphological Completion (38 items). The three supplemental subtests include Word Discrimination (28 items), Phonemic Analysis (22 items), and Word Articulation (25 items).[1] Administration time is 35 to 50 minutes for core subtests and an additional 30 minutes to administer all supplemental subtests (which should be administered during a separate testing period). Ceilings are established for all core subtests when the student responds incorrectly to five consecutive items (there are no basals). All items in the supplemental subtests should be administered. The authors note that ceiling effects were found for the three subtests measuring grammar at the 8-year-old level, but they did not include more difficult items because the TOLD-P:4 is used primarily to identify deficiencies in oral language rather than to distinguish proficiency of highly skilled speakers.

Other Languages: None.

Uses of Information: The authors state that the TOLD-P:4 may be used to identify students who are below level in language proficiency; to determine strengths and weaknesses in language skills; to monitor progress in language as a result of an intervention; and to assess language for research purposes. As mentioned under the Description, the authors note that ceiling effects were found for the three subtests measuring grammar at the 8-year-old level, further underscoring the usefulness of the assessments to identify deficiencies of basic skills as opposed to competences of higher language skills.

Methods of Scoring: An assessor who has thoroughly reviewed the manual and examiner record booklet should score the test. Raw scores for each subtest are calculated by summing correct responses. Raw scores may be converted to age equivalents,[2] percentile ranks, and scaled scores in which scores are expressed on a scale with a mean of 10 and standard deviation of 3. Using the core subtest scaled scores, composite scales (Listening, Organizing, Speaking, Grammar, Semantics, and Spoken Language) on language competence may be calculated in order to distinguish language competence from speech competence (measured by the supplemental subtests on phonology). Composite scores are expressed on a scale with a mean of 100 and standard deviation of 15. A single subtest may factor into more than one composite index.

Interpretability: The manual provides information on how to interpret scores and describes the subtests. The appendix contains conversion tables. Only assessors with strong clinical skills should interpret scores on the TOLD-P:4, and the authors caution assessors against relying solely on results to diagnose spoken language difficulties.

Reliability:[3]
(1) Internal consistency reliability: Coefficients across age groups ranged from 0.80 to 0.97 for core subtests, from 0.83 to 0.97 for supplemental subtests, and from 0.87 to 0.97 for composite scores. The entire normative sample was used to estimate the level of reliability for scores by subtest.
(2) Test-retest reliability: The correlations of scores between two administrations (one- to two-week intervals) ranged from 0.81 to 0.87 for core subtests, from 0.78 to 0.84 for supplemental subtests, and from 0.84 to 0.92 for composite scores. A sample of 89 students (age 4 through 8 years) from Austin, Texas, was used for test-retest reliability.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: Two PRO-ED staff independently scored the same 50 tests from the normative sample. They did not have a background in language arts or assessment but were familiar with the manual and examiner record booklet. The correlations between their sets of scores ranged from 0.97 to 0.99 across all subtests and composite scores.

Validity Evidence:
Construct/Concurrent validity: Item statistics were examined to improve the level of reliability for scores and demonstrate evidence of construct validity. Item discrimination below 0.30 and items with p-values outside the range of 0.15 to 0.85 were dropped after pilot testing. Remaining items were ordered according to difficulty level. New items were created for each subtest of the current version of the TOLD-P by following the same process. The subtest median item discrimination coefficients averaged across age groups ranged from 0.33 to 0.62 for the nine subtests, and the subtest median percentages of difficulty averaged across age groups ranged from 41 to 82 percent. Two factors were identified in the factor analysis, General Oral Language (all six of the core subtests loaded on this factor) and Word Articulation (the three supplemental subtests loaded on this factor). The normative sample was used for these analyses.

TOLD-P:4 scores were correlated with the following measures of spoken language: the Pragmatic Language Observation Scale (PLOS), the Test of Language Development-Intermediate: Fourth Edition (TOLD-I:4), and the Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV). Correlation coefficients between the TOLD-P:4 and the PLOS total standard score ranged from 0.34 to 0.63 for core subtests, from 0.40 to 0.61 for supplemental subtests, and from 0.50 to 0.64 for composite scores. Correlation coefficients between the TOLD-P:4 and the TOLD-I:4 Spoken Language composite ranged from 0.31 to 0.48 for core subtests and from 0.42 to 0.55 for composite scores. Correlation coefficients between the TOLD-P:4 and the WISC-IV Verbal Comprehension composite ranged from 0.30 to 0.66 for subtests, and correlations for most composite scores ranged from 0.56 to 0.76. The exceptions were Grammar ($r = 0.09$) and Organizing ($r = 0.12$). Correlations between the TOLD-P:4 with the TOLD-I:4 and with the WISC-IV for supplemental subtests were not given because supplemental subtests are not appropriate for students over age 7 (i.e., TOLD-I:4 and WISC-IV are normed for age 6 through 17 years). Correlation coefficients corrected for restricted range and attenuation for measurement error were also presented. Samples ranged from 31 to 663 students age 4 through 8 years from across the nation.

The authors present positive predictive values and percent agreement relating to the ability of the TOLD-P:4 to detect language problems of students assessed on other measures. The measures included the PLOS ($N = 663$), TOLD-I:4 ($N = 71$), and Global Spoken Language.[4] Sensitivity

indices ranged from 0.74 to 0.75 and specificity indices from 0.87 to 0.88. Positive predictive values ranged from 0.70 to 0.71, and percent agreement from 83 to 85 percent across the three measures.

Scores on the TOLD-P:4 are expected to increase with age. Mean raw scores were provided for each of the five age categories (4 through 8 years) across subtests. Mean raw scores ranged from 9 (age 4) to 27 (age 8) across all subtests for the different ages. Correlations were noted between age and the Relational Vocabulary core subtest ($r = 0.58$) and the Word Discrimination and Word Articulation supplemental subtests ($r = 0.40$). Information on the sample was not provided.

Based on the norming sample, students' standard scores and composite scores by exceptionality status (i.e., gifted and talented individuals and individuals with attention-deficit/hyperactivity disorder, learning disabilities, and language disorders) fell within the expected range of "average" (standard scores of 8 through 12; composite scores of 90 through 110) and "below average" (standard scores of 6 and 7; composite scores of 80 through 89). Many of the subtests and composite scores for individuals identified as gifted and talented were "above average" (standard scores of 13 and 15; composite scores of 111 through 120).

Internal consistency reliability was also provided for the exceptionality groups based on the norming sample. Reliability estimates for gifted and talented students ranged from 0.80 to 0.95 for core subtests and from 0.88 to 0.92 for supplemental subtests. Test-retest reliability coefficients ranged from 0.90 to 0.98 for composite scores. Reliability estimates for students with attention-deficit/hyperactive disorder ranged from 0.82 to 0.95 for core subtests and from 0.85 to 0.92 for supplemental subtests. Test-retest reliability coefficients ranged from 0.92 to 0.98 for composite scores. Reliability estimates for students with learning disabilities ranged from 0.88 to 0.94 for core subtests and from 0.84 to 0.87 for supplemental subtests. Test-retest reliability coefficients ranged from 0.94 to 0.99 for composite scores.

Predictive validity: No information available.

Bias Analysis: A logistic regression procedure was used to detect differential item functioning (DIF) among the subgroups in the norming sample. Comparisons were conducted for gender and race/ethnicity (i.e., Black versus non-Black students, Hispanic versus non-Hispanic students).

The authors do not state whether the seven items showing DIF (three for gender and four for race/ethnicity) were removed from the assessment but assert it is non-biased with regard to gender, race, and ethnicity. In addition, the authors present standard scores by gender and race/ethnicity in which each group performed at the average level with scores ranging from 8 to 12 points for the subtests and 90 to 110 points for the composites.

Training Support: The assessor should be familiar with the manual and examiner record booklet and should conduct several practice administrations.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: None.

Previous Version: As compared to the third edition of the TOLD-P, the newest version does not present floor effects; ceiling effects have been addressed. In addition, item bias was further explored, and validity studies on sensitivity and specificity were provided. Directions for each subtest have been added to the examiner record booklet, and the manual has been revised to be more user-friendly.

NCEE or REL Study Use:[5] Even Start Classroom Literacy Interventions and Outcomes (CLIO) Study

[1] Supplemental subtests should not be administered to students over age 7 unless they exhibit problems in the corresponding skill areas.

[2] The authors recommend caution in using age equivalents.

[3] The *Critical Reviews and Research Findings for TOLD-P: 1977–2007* provides additional information on the reliability and validity research of the TOLD-P over the years.

[4] Global Spoken Language is a meta-variable made up of the PLOS and TOLD-I:4 scores across both samples.

[5] See Table F.1 for web address.

References:

Newcomer, Phyllis L., and Donald D. Hammill. *Critical Reviews and Research Findings for TOLD-P: 1997–2007*. Austin, TX: PRO-ED, 2008.

Newcomer, Phyllis L., and Donald D. Hammill. *Test of Language Development-Primary: Fourth Edition*. *Examiner's Manual*. Austin, TX: PRO-ED, 2008.

TEST OF PRESCHOOL EARLY LITERACY (TOPEL), 2007

Authors: Christopher J. Lonigan, Richard K. Wagner, Joseph K. Torgesen, and Carol A. Rashotte		**Type of Assessment:** Individually administered adaptive assessment **Domain:** Reading (print concepts, letter recognition and naming, and phonological awareness) and language arts/language proficiency (vocabulary)
Publisher: PRO-ED Inc. 800-897-3202 http://www.proedinc.com		**Grade/Age Range:** 3 through 5 years **Administration Interval:** None described
Material, Training, and Scoring Costs: Complete kit (Examiner's Manual, Picture Book, and 25 Record Booklets in a storage box): $214		**Personnel and Training Requirements** **Credentials Required for Use:** No special qualifications required/noted **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours) The assessor should have knowledge and experience in test administration, test scoring, and interpretation of norm-referenced results.
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: The norming sample consisted of 842 students, including 212 3-year-olds, 313 4-year-olds, and 317 5-year-olds, from 12 states tested in 2004. This convenience sample was based on assessors in the PRO-ED customer files who tested 20 students each. The norming sample closely approximates the U.S. population, based on the 2001 Bureau of the Census, for region, gender, race, Hispanic ethnicity, family income, parent education attainment, and exceptionality status (such as a language disorder, attention-deficit/hyperactivity disorder, or a disability). Developers present age-stratified demographic variables that parallel national, school-age-specific estimates.		**Summary** **Initial Material Cost:** 3 ($200 to $500) **Time to Administer:** 30 minutes **Ease of Administration and Scoring:** 3 (administered and scored by highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 3 (normed within past 10 years and nationally representative)

NARRATIVE

Description: The TOPEL is an individually administered adaptive assessment of early literacy normed for 3- through 5-year-olds. The assessment consists of 98 items with three subtests: Print Knowledge (36 items), Definitional Vocabulary (35 items), and Phonological Awareness (27 items). The Print Knowledge subtest measures written language conventions and alphabet knowledge. The student points to, identifies, or says the sounds associated with letters, words, and aspects of print. The Definitional Vocabulary subtest measures a student's single word oral vocabulary and definitional vocabulary. The student identifies a picture and answers a question about the picture's attributes. The Phonological Awareness subtest measures elision and blending abilities. The student says words after being instructed to drop sounds (elision) and combines separate sounds into a word after listening to the sounds (blending). The TOPEL takes approximately 30 minutes to administer. Each subtest contains item sets, which are groups of items assessing the same skill. The Print Knowledge and Phonological Awareness subtests contain multiple item sets, whereas the Definitional Vocabulary subtest contains one item set. The assessor administers all item sets within each subtest. All three subtests have a ceiling rule of three consecutive incorrect responses, which are applied to each item set within each subtest.

Other Languages: The precursor to the TOPEL, the Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP) (Lonigan et al. 2002; Lonigan 2002), includes a Spanish language version that has been used in some studies. No psychometric data, particularly about equivalence, are available for the Spanish version of the Pre-CTOPPP.

Uses of Information: The TOPEL is used to quantify and measure change over time in literacy-related abilities. The developers note that the assessment may also be used to (1) identify students at risk of having or developing literacy-related problems and (2) monitor progress in early literacy-related skills in response to an intervention or program.

Methods of Scoring: Assessors code each correct response as "1" and each incorrect response as "0." Raw scores reflect the total correct responses in all item sets to the last item in the ceiling. A total composite score (Composite Early Literacy Index) and subtest scores are computed. The manual includes appendices with conversions of raw scores into standard scores and percentile ranks.

Interpretability: The manual includes extensive instructions for interpreting below average, average, and above average standard scores for the subtests and the composite measure as well as general information on what standard scores mean. Developers indicate that standard scores provide the clearest indication of a student's performance on the TOPEL. The manual briefly discusses interpretations of raw scores and percentile ranks.

Reliability:
(1) Internal consistency reliability: Cronbach's alphas were calculated for three age groups (3, 4, or 5 years) for composite and subtest scores. For the composite score, coefficients ranged across age groups from 0.95 to 0.96 and from 0.93 to 0.96 for Print Knowledge subtest scores, 0.94 to 0.95 for Definitional Vocabulary, and 0.86 to 0.88 for Phonological Awareness.
(2) Test-retest reliability: The sample consisted of 45 3- to 5-year-olds from Mandan, North Dakota, who were primarily White and female. Test-retest reliability (with a two-week interval)

of standard scores ranged from 0.81 to 0.91. The authors noted without elaboration a practice effect for the Phonological Awareness subtest.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: Two trained assessors independently scored 30 randomly selected protocols from the normative sample. Reliability coefficients using standard scores for subtests and the composite ranged from 0.96 to 0.98.

Validity Evidence:
TOPEL developers noted that subtests for Print Knowledge and Phonological Awareness were based on their research over the past decade. The Definitional Vocabulary subtest contained frequently used word items from several sources, such as word frequency guides, works of literature, popular fiction and non-fiction used in schools, and early vocabulary lists and analyses. Developers describe various field tests with preschool-age children, generally from Florida, using an iterative process to reduce the pool of items. Items were removed or modified based on inconsistency of students' response patterns, item difficulty level, or low correlations between items and total scores.

Construct/Concurrent validity: Developers analyzed item validity and item difficulty of finalized subtests on the full normative sample. Median item discrimination coefficients ranged from 0.38 to 0.66. Median item difficulty, which reflects the percentage of students who passed a given item, ranged from 0.20 to 0.84.

Scores on the three TOPEL subtests were correlated with scores on the Test of Early Reading Ability-Third Edition (TERA-3) Alphabet subtest, the TERA-3 Reading Quotient, the Expressive One-Word Picture Vocabulary Test-2000 Edition (EOWPVT), the Get Ready to Read! Screening Tool, and the Comprehensive Test of Phonological Processing (CTOPP) Elision Blending Words subtests. The sample consisted of 154 3- to 5-year-olds from Tallahassee, Florida, of whom the majority was male (60 percent) and White (89 percent). Uncorrected correlations between the TOPEL Composite Early Literacy Index and the TERA-3 Reading Quotient and the Get Ready to Read! Screening Tool were 0.63 and 0.60, respectively. The TOPEL Print Knowledge subtest scores correlated 0.74 with the TERA-3 Alphabet scores. The TOPEL Definitional Vocabulary subtest scores correlated 0.62 with the EOWPVT scores. The TOPEL Phonological Awareness subtest scores correlated 0.52 and 0.55 with the CTOPP Elision and the Blending of Words scores, respectively. In addition, the three individual TOPEL subtests were correlated with the TERA-3 Reading Quotient and Get Ready to Read!, with uncorrected coefficients ranging from 0.37 to 0.57. Developers also provided corrected correlations to account for the effects of range.

With respect to subgroup differences, the authors examined chronological age and Hispanic American-bilingual status in relation to TOPEL performance for the entire normative sample. Chronological age was positively related to TOPEL performance on the three subtests such that raw score means increased with age. Correlation coefficients between the means of raw scores for the three age groups and each subtest were 0.49, 0.54, and 0.56 for Phonological Awareness, Print Knowledge, and Definitional Vocabulary, respectively. Hispanic American-bilingual students demonstrated standard scores below the average range (i.e., 90 to 110) for Definitional Vocabulary, Phonological Awareness, and the Composite score (mean = 82, 89, 84, respectively) but average scores for Print Knowledge (mean = 92). Developers noted that below average scores for Hispanic American-bilingual students support the validity of the assessment.

Predictive validity: No information available.

Bias Analysis: Three types of analyses were conducted to examine the impact on various groups of test takers: (1) Differential Item Functioning (DIF), (2) comparison of mean standard scores, and (3) internal consistency coefficients. DIF analysis was conducted on the entire normative sample of 3- to 5-year-olds, and the groups compared included gender, race (Black versus non-Black), and ethnicity (Hispanic versus non-Hispanic). Developers neither reported the groups favored in each item comparison nor removed any items based on DIF analyses, but they reported several other findings. In the DIF by gender analysis, one item in the Definitional Vocabulary subtest had a moderate effect size. In the DIF by race analysis, two items, including one in the Print Knowledge subtest and one in the Definitional Vocabulary subtest, had moderate effect sizes. In the DIF by ethnicity analysis, four items in the Definitional Vocabulary subtest had moderate or large effect sizes. The mean standard scores for gender (male, female) and race/ethnicity (White, Black, and Hispanic-English-only students) were average, with standard scores ranging from 92 to 105. Cronbach's alphas were calculated for the composite and each subtest score by subgroup (male, female, White, Black, and Hispanic) and ranged from 0.85 to 0.97.

Training Support: The manual provides information on the basic administration of the assessment.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: None.

Previous Version: None.

NCEE or REL Study Use:[1] The Pre-CTOPPP, a precursor to the TOPEL, was used in the Even Start Classroom Literacy Interventions and Outcomes (CLIO) Study and the National Evaluation of Early Reading First.

[1] See Table F.1 for web address.

References:

Lonigan, Christopher J. "Pre-CTOPP Subtest Statistics for Preschool Comprehensive Test of Phonological and Print Processing by Age Group." Available at [http://www.psy.fsu.edu/~lonigan/data.pdf]. 2002.

Lonigan, Christopher J., Richard K. Wagner, Joseph K. Torgesen, and Carol A. Rashotte. *Preschool Comprehensive Test of Phonological & Print Processing (Pre-CTOPPP)*. Florida State University, Department of Psychology, 2002.

Lonigan, Christopher J., Richard K. Wagner, Joseph K. Torgesen, and Carol A. Rashotte. *TOPEL: Test of Preschool Early Literacy*. Austin, TX: PRO-ED, 2007.

TEST OF SILENT CONTEXTUAL READING FLUENCY (TOSCRF), 2006

Authors: Donald D. Hammill, J. Lee Wiederholt, and Elizabeth A. Allen		**Type of Assessment:** Group-administered or individual assessment **Domain:** Reading (word recognition; reading comprehension)
Publisher: PRO-ED Inc. 800-897-3202 http://www.proedinc.com		**Grade/Age Range:** 7 years through 18 years, 11 months **Administration Interval:** Every 2 months
Material, Training, and Scoring Costs: TOSCRF kit (includes assessor's manual and 25 student record forms for each alternate form): $207		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2 (bachelor's degree with coursework in measurement and domain) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours) Assessors should review the manual carefully and practice administering and scoring the test until they accurately score the 10 practice tests provided.
Languages: English		**Alternate Forms:** Four forms; administration interval no more than every two months
Representativeness of Norming Sample: The norming sample consisted of 1,898 students ages 7 to 18 years from 23 states in cities and rural areas tested in 2004. With respect to geographic region, gender, family income, parent education level, special education status, and age, the sample approximates the United States according to the Bureau of the Census's *The Statistical Abstract of the United States, 2002*. The sample was also stratified by age, across geographic region, ethnicity, Hispanic status, gender, family income, and parent education.[1]		**Summary** **Initial Material Cost:** 3 ($200 to $500) **Time to Administer:** 10 minutes (3 minutes testing time) **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 1[2] (none described) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 2 (older than 10 years or not nationally representative)

NARRATIVE

Description: The Test of Silent Contextual Reading Fluency (TOSCRF) is administered to either groups or individuals to assess their silent reading ability of English. It is appropriate for students between ages 7 years and 18 years, 11 months. Students receive a booklet of word passages with no spaces, punctuation, or other sentence breaks and are asked to mark a line between each appropriate word given the context of the passage. The passages increase in length and difficulty as the students progress through the assessment. Students complete as many passages as possible in 3 minutes; the total estimated administration time, including instructions and practice items, is 10 minutes. The authors note no ceiling effects but do note floor effects at the 7-year-old level; that is, the TOSCRF may be used to identify students with reading difficulties but cannot assess the degree of difficulty for 7-year-olds.

Other Languages: None.

Uses of Information: The TOSCRF is used to measure a student's silent reading skills. The authors note that the assessment may also be used to screen for students struggling with reading as well as for above-average readers; measure a student's silent contextual reading fluency rather than word-level reading fluency; monitor reading skill progress; measure intervention effectiveness; and validate other reading instruments. The authors also note that the assessment can estimate the degree of reading difficulty, except for 7-year-olds as noted above.

Methods of Scoring: The assessor begins scoring from the last line the student completed, scoring backwards until he/she reaches the point where the student has correctly identified all the words in one passage or all words have been scored. Students receive credit (one point per word) for each word correctly identified, the sum of which is the student's raw score. In determining correctness, several additional rules pertain to how to score the marks based on angle, size, and placement. Reviewers have noted that scoring can be complex (Smith 2007; Soares 2007). A student's raw score is used to calculate normative scores such as standard scores, percentiles, and age and grade equivalents. One chart converts raw scores into standard scores and percentiles; separate charts convert raw scores into "reading age" equivalents and grade equivalents.

Interpretability: Only persons with formal training in psychological testing and statistics should interpret the results of the TOSCRF. Each student record form includes a section for the assessor to record his/her interpretation of the normative scores as well as recommendations for further assessment. In an attempt to help readers understand standard scoring, the manual uses IQ scores to illustrate the meaning of a mean of 100 and a standard deviation of 15. One reviewer notes that the reference to IQ scores may confuse naïve users about the meaning of the scores by suggesting a relationship with IQ scores (Smith 2007). The reviewer also notes that caution should be used in interpreting scores for students residing in any of the areas of the United States not represented in the norming sample (Smith 2007). In addition, the authors caution about interpretations of age and grade equivalent scores, as interpolation, extrapolation, and smoothing were used in the scores' derivation.

Reliability:
(1) Internal consistency reliability: No information available.

(2) Test-retest reliability: Across the four forms, correlation coefficients (with about a two-week interval between administrations) ranged from 0.83 to 0.92 for elementary students; from 0.69 to 0.79 for middle school students; and from 0.93 to 0.97 for high school students. For the entire combined sample, coefficients ranged from 0.83 to 0.89 across the forms. (Authors note that second testing means were "appreciably higher" than the first testing, suggesting practice effects.)
(3) Alternate form reliability: Correlations among all possible two-form combinations using the four alternate forms with immediate administration ranged from 0.76 to 0.90 across ages 7 to 18 years. The averaged correlation coefficients between forms ranged from 0.82 to 0.86 for all ages; averaged correlations for all forms across each age, 7 to 18 years, ranged from 0.82 to 0.88. The individual correlations among all possible two-form combinations using the four alternate forms with about a two-week interval between administrations ranged from 0.80 to 0.89.
(4) Inter-rater reliability: Correlation coefficients across four forms using four raters were all 0.99.

Validity Evidence:
The authors liken the assessment to hidden word search puzzles and note that such word-string tests have been previously used in practice. Passages were chosen from pre-existing assessments, including the Gray Oral Reading Tests-Fourth Edition (GORT-4) and the Gray Silent Reading Tests. The manual describes how sentences were constructed for the specific test format.

Construct/Concurrent validity: The authors compared the TOSCRF (including all four forms) against five assessments: the GORT-4; the Stanford Achievement Test Series-Ninth Edition Total Reading score (Stanford 9); the Test of Silent Word Reading Fluency (TOSWRF); the Test of Word Reading Efficiency (TOWRE); and the Woodcock-Johnson III (WJ-III). The authors also constructed a "Global Reading" score from the five assessments for analysis. Samples totaled about 300 students, with the majority male with special needs. Only the TOSWRF and the TOWRE were collected concurrently with the TOSCRF. All other scores came from archival sources. The authors calculated the difference in standard score means between the TOSCRF and each of the five assessments and composite score; the differences were significant for two of the archival scores (GORT-4 Total Score and the WJ-III Broad Reading score) but insignificant for the Stanford 9 Total Reading score, the TOSWRF Total Score, the TOWRE Total Quotient, and the "Global Reading" score. The authors examined the relationship of the TOSCRF to these measures and found that the measures administered concurrently had stronger bivariate relationships than did the archival test scores with the TOSCRF: the GORT-4 Total Score (0.45 to 0.52), the Stanford 9 Total Reading score (0.45 to 0.54), the WJ-III Broad Reading score (0.55 to 0.66), the TOSWRF Total Score (0.71 to 0.79), the TOWRE Total Quotient (0.80 to 0.86), and the "Global Reading" score (0.69 to 0.73).

The authors correlated the four forms of the TOSCRF with the following assessments: the Stanford 9 Total Math (0.36 to 0.45) and Vocabulary (0.33 to 0.40) and the WJ-III Calculation (0.29 to 0.40), Spelling (0.58 to 0.70), and Academic Skills (0.52 to 0.57). Finally, the authors correlated scores on the TOSCRF with scores on the Wechsler Intelligence Scale for Children (WISC), Third and Fourth editions. The correlations ranged from 0.30 to 0.37 on the WISC Verbal Scale, from 0.26 to 0.30 on the WISC Performance Scale, and from 0.34 to 0.40 on the WISC Full Scale.

Additionally, the authors determined the ability of the TOSCRF to identify students with reading problems with use of the GORT-4 (N = 119), the Stanford 9 (N = 103), the TOSWRF (N = 275), the TOWRE (N = 42), and the WJ-III (N = 130) as well as with the authors' constructed Global Reading score (N = 641). The sensitivity index ranged from 0.71 to 0.81, the specificity index ranged from 0.59 to 0.87, and the positive predictive value ranged from 0.46 to 0.84. No information on the age of the students in the sample was provided.

Predictive validity: No information available.

Bias Analysis: The authors reported immediate-administration alternate form reliability correlations for 11 subgroups from the norming sample: males (N = 991; 0.82 to 0.88), females (N = 906; 0.81 to 0.84), White students (N = 1,530; 0.83 to 0.86), Black (N = 231; 0.73 to 0.85), Asian/Pacific Islander (N = 80; 0.82 to 0.89), Hispanic (N = 214; 0.83 to 0.85), gifted and talented (N = 70; 0.78 to 0.81), students with attention-deficit/hyperactivity disorder (N = 134; 0.83 to 0.87), a learning disability (N = 154; 0.83 to 0.89), deaf/hard-of-hearing (N = 49; 0.86 to 0.94), and poor readers (N = 316; 0.81 to 0.86).

Training Support: The manual includes 10 practice tests for achieving scoring mastery. The authors otherwise suggest enrollment in college-level courses on assessment or in workshops or in-service training provided by local schools or private consultants.

Adaptations/Special Instructions for Individuals with Disabilities: If during a group administration a student is unable to identify the sample words or exhibits problems using a pen or pencil, the authors note that the assessment should not be scored.

Alternate Forms: TOSCRF has four parallel forms—Forms A through D. Authors recommend re-testing students with alternate forms every 2 months at most.

Previous Version: None.

NCEE or REL Study Use:[3] Evaluation of Reading Comprehension Programs

[1] Smith (2007) questioned representativeness by geographic region, noting an under-representation of the Southeast, Southwest, and Mountain states.

[2] The reliability rating refers to internal consistency reliability, which was required for direct assessments (see Appendix A). See the reliability section in the profile narrative for information available on test-retest and alternate form reliability.

[3] See Table F.1 for web address.

References:

Hammill, Donald D., J.L. Wiederholt, and Elizabeth A. Allen. *Test of Silent Contextual Reading Fluency: Examiner's Manual*. Austin, TX: PRO-ED Inc., 2006.

Smith, Lisa F. “Review of the Test of Silent Contextual Reading Fluency.” In *The Seventeenth Mental Measurements Yearbook,* edited by Kurt F. Geisinger, Robert A. Spies, Janet F. Carlson, and Barbara S. Plake. Lincoln, NE: Buros Institute of Mental Measurements, 2007.

Soares, Louise M. “Review of the Test of Silent Contextual Reading Fluency.” In *The Seventeenth Mental Measurements Yearbook,* edited by Kurt F. Geisinger, Robert A. Spies, Janet F. Carlson, and Barbara S. Plake. Lincoln, NE: Buros Institute of Mental Measurements, 2007.

TEST OF SILENT WORD READING FLUENCY (TOSWRF), 2004

Authors: Nancy Mather, Donald D. Hammill, Elizabeth A. Allen, and Rhia Roberts

Publisher:
PRO-ED, Inc.
800-897-3202
http://www.proedinc.com

Material, Training, and Scoring Costs:
TOSWRF Test Kit (Examiner's Manual, 50 Student Record Forms A, 50 Student Record Forms B, storage box): $147

Languages: English

Representativeness of Norming Sample:
The norming data were collected in spring and fall 2001 and spring and summer 2002. The sample included 3,592 students in 32 states in 4 large regions of the United States. The authors state that the sample approximated the U.S. population (as reflected in 2001 Census data) with respect to region, gender, race, ethnicity, parental education level, disability status, and age. White and Black students and students from other racial/ethnic backgrounds made up 77, 12, and 11 percent of the sample, respectively; 10 percent of the sample was Hispanic. Students were 6- through 17-years-old, with males and females equally represented. Nineteen percent of the students were diagnosed with special needs due to disability or giftedness. Sample demographic characteristics also approximated national Census data when stratified by age.

Type of Assessment: Individual or group-administered assessment
Domain: Reading (fluency; reading comprehension)

Grade/Age Range: 6 years, 6 months through 17 years, 11 months
Administration Interval: No minimum interval; both forms may be used in a single test administration

Personnel and Training Requirements
Credentials Required for Use: Level 2 (bachelor's degree with coursework in measurement and domain)
Personnel for Administration: Individual with basic clerical skills with some training
Training for Administration: Minimal (1 to 2 hours)
Assessors should be able to score 10 practice test forms correctly in the manual before scoring an actual test.

Alternate Forms: Two forms; no minimum administration interval

Summary
Initial Material Cost: 2 ($100 to $200)
Time to Administer: 3 minutes for single form or 10 minutes for both forms
Ease of Administration and Scoring: 3 (administered and scored by a highly trained individual)
Reliability: 1[1] (none described)
Predictive Validity: Not available
Construct/Concurrent Validity: Available
Norming Sample Characteristics: 3 (normed within past 10 years and nationally representative)

NARRATIVE

Description: The TOSWRF is an individual or group-administered paper-and-pencil assessment that measures students' reading fluency as demonstrated by students' ability to recognize printed words accurately and efficiently. The assessment presents the words in 32 rows with no spaces between words, and students are asked to identify printed words by drawing lines between the boundaries of words. The rows of words are listed in ascending order of reading difficulty. Students are asked to identify as many words as possible in 3 minutes. The authors state that the TOSWRF measures word comprehension in addition to reading fluency. The assessment may be used with students age 6 years, 6 months through 17 years, 11 months, with one or both of two equivalent forms (A and B). The administration time for one form is 3 minutes; both forms may be administered in 10 minutes.

Other Languages: None.

Uses of Information: The TOSWRF may be used as an individual or group screening measure to identify students with reading difficulties. They also note that results from the assessment may be used to estimate general reading ability and to identify poor readers. The information provided by the measure should not, however, be used as the sole basis for eligibility or placement decisions. The measure may also be used to monitor the progress of individual students or groups of students. Researchers and program evaluators may use it to assess students' word reading fluency over time or as a student outcome variable in studies comparing the effectiveness of different instructional settings. It may also be used for sample selection purposes and to validate other reading measures.

Methods of Scoring: To score the TOSWRF, the assessor reviews the words identified by the student on the answer sheet, beginning with the last row attempted by the student. Working backward, the assessor identifies the point at which the student correctly identified all words on two consecutive rows or until all words have been scored. From that point to the beginning of the assessment, the assessor awards one point for each possible word (whether or not correctly identified). The assessor also awards one point for each correctly identified word in and after the two consecutive rows of correct identifications. The manual includes instructions for scoring skipped rows, misplaced lines, and other irregularities and includes a scoring key in the appendix. At the point at which the student does not identify all the words correctly for two consecutive rows, the assessor computes the total raw score as the sum of all correctly identified words. Using norm tables, the assessor converts raw scores into standard scores, percentile ranks, and age and grade equivalents. To maximize the reliability of the TOSWRF scores, the assessor may convert the sum of the standard scores for Form A and Form B into a composite standard score.

Interpretability: The manual includes seven descriptive ratings for interpreting standard scores, ranging from "very superior" to "very poor." Given that the TOSWRF assesses many aspects of reading (word identification and speed, word comprehension), the authors consider it a valid instrument for screening students with reading difficulties and for assessing general reading ability.

Reliability:
(1) Internal consistency reliability: Given that the TOSWRF is a timed assessment without distinct items, the authors stated that it is not possible to estimate internal consistency reliability properly for the measure's scores. Young (2005) noted that, while the TOSWRF is a timed measure, internal consistency reliability could be calculated on the number of words correctly identified for each line, considering each line as an item.
(2) Test-retest reliability: The authors calculated test-retest reliability coefficients with five samples of students in West Virginia, Maryland, and Kansas (total N = 200). They collected data from one sample of elementary students (7- to 10-year-olds), two samples of middle school students (12- to 15-year-olds), and two samples of high school students (15- to 17-year-olds). Almost all of the students in the sample were White, non-Hispanic, and typically developing. Both forms (A and B) were administered twice to all of the students with an interval of about two weeks between administrations. The authors correlated the two sets of Form A scores with each other, and the two sets of Form B scores with each other. Across all of the samples, the mean test-retest correlations ranged from 0.45 to 0.82; the correlations for both high school samples were below 0.70. Across samples and forms, the average test-retest coefficient was 0.69. The authors also report and interpret correlation coefficients that are corrected because of restriction of range.
(3) Alternate form reliability: Students in the norming sample completed both Forms A and B of the TOSWRF in one test session (half completed Form A followed by Form B; the other half completed Form B followed by Form A). The authors correlated norming sample standard scores for Forms A and B at 12 age intervals (age 6 through 17 years). Alternate form correlation coefficients across age intervals ranged from 0.77 to 0.91, with an average coefficient (calculated by the *z*-transformation method for averaging correlation coefficients) of 0.86. The authors also calculated alternate form (delayed administration) reliability coefficients with the five samples of students described above (see Test-retest reliability). Across samples, the average alternate form (delayed administration) reliability coefficients ranged from 0.36 to 0.79 (four of six were below 0.70). The authors also report and interpret correlation coefficients that are corrected because of restriction of range.
(4) Inter-rater reliability: One of the co-authors scored 486 TOSWRF protocols drawn from the validity samples (described below), and 10 colleagues independently scored a subset of the same protocols. The students were 6- to 17-year-olds, lived in seven states, and had a broad range of reading competency. Inter-rater reliability coefficients were 0.99 for Form A and 0.99 for Form B.

Validity Evidence:
The design of the TOSWRF was informed by Guilford's Structure of Intellect model (Guilford and Hoepfner 1971) that used word search tasks to assess cognitive abilities as well as by subsequent measures that used timed word find or "word-strings-without-spaces" formats to assess speed of word recognition (Meeker and Meeker 1975; Miller-Guron 1996). Mather et al. (2004) built on Miller-Guron's (1996) Wordchains measure by ordering the words by increasing difficulty as determined by a graded word frequency list (Taylor et al. 1989).

Construct/Concurrent validity: All coefficients presented are uncorrected, but the authors also report and interpret correlation coefficients that are corrected because of restriction of range. Mather et al. (2004) conducted validity studies correlating TOSWRF scores with scores on other assessments of reading fluency, word identification, and comprehension. They collected data

from four samples, one of which consisted mostly of students with disabilities (85 percent). The authors correlated TOSWRF scores with scores on Wordchains (Miller-Guron 1999) and the Sight Word Efficiency and Phonemic Decoding Efficiency subtests from the Test of Word Reading Efficiency (TOWRE). Correlations between scores on Forms A and B of the TOSWRF and Wordchains were 0.71 and 0.67, respectively, and correlations between both TOSWRF forms and the TOWRE subtest scores ranged from 0.73 to 0.78. They also correlated scores on both TOSWRF forms with scores from the Word Identification subtest of the Woodcock Reading Mastery Test-Revised, Normative Update (WRMT-R/NU) and the Letter-Word Identification and Passage Comprehension subtests of the Woodcock-Johnson Psycho-Educational Battery-Revised Tests of Achievement (WJ-R). Correlation coefficients ranged from 0.47 to 0.53. Mather et al. (2004) also correlated TOSWRF scores with scores on the WJ-R Passage Comprehension subtest. Correlations ranged from 0.33 to 0.61. Mee Bell et al. (2006) reported correlations between scores on the TOSWRF and four subtests of the Woodcock-Johnson III Tests of Achievement (WJ III ACH)—Letter Word Identification, Reading Fluency, Passage Comprehension, Spelling, and Broad Reading Cluster—ranging from 0.58 to 0.66.

Researchers have also correlated TOSWRF scores with other types of achievement scores and ability test scores. Mather et al. (2004) reported correlations ranging from 0.24 to 0.69 between TOSWRF scores and scores from WJ-R Reading, Math, and Broad Knowledge Skills. Correlations between TOSWRF scores and Verbal, Performance, and Full Scale scores on the Wechsler Intelligence Scale for Children--Third Edition (WISC-III) ranged from 0.26 to 0.33. In a study of elementary school students with reading and/or other learning difficulties, Mee Bell et al. (2006) reported that TOSWRF scores correlated with Comprehensive Test of Basic Skills (CTBS) Spelling scores ($r = 0.60$), but not with CTBS Vocabulary, Reading, Reading Composite, or Language Composite scores.

The authors reported evidence that students' performance on the TOSWRF differed by age, exceptionality category, and reading competency. TOSWRF scores increased with age across 12 age intervals, and age scores on Forms A and B correlated 0.77 and 0.76, respectively. Gifted students exhibited higher-than-average mean standard scores, and students with disabilities had below average standard scores. TOSWRF scores also discriminated between students identified as poor readers (according to scores on the TOWRE and Wordchains). Mather et al. (2004) reported sensitivity indices ranging from 0.62 to 0.80, specificity indices from 0.91 to 0.93, positive predictive values of 0.70 to 0.75, and percent agreement rates ranging from 84 to 89 percent between scores on the two measures. They point to these results as evidence that the TOSWRF may be used to screen students with general reading problems.

Predictive validity: No information available.

Bias Analysis: The authors compared standard score means and standard deviations for the total sample, males and females, and selected racial/ethnic subgroups. For Form A, the standard score means for males and females were 99 and 102, respectively (similar means were obtained for Form B). They reported some variation in Form A mean scores among ethnic groups (ranging from 90 for American Indian/Eskimo/Aleut students to 109 for Asian American/Pacific Islander students (Form B mean scores were identical). They concluded that, despite the variation, mean scores for all groups were within the average range and "within expectations."

Training Support: The manual includes 10 practice test forms so that trainees can practice and master scoring the test before they administer it.

Adaptations/Special Instructions for Individuals with Disabilities: The authors do not specify adaptations or special instructions for individuals with disabilities but do caution that the assessment should not be administered to students with eye-hand coordination difficulties.

Alternate Forms: Equivalent Forms A and B are available. There is no minimum time interval between administrations. To increase reliability, assessors may administer both forms in one test session and compute a composite score.

Previous Version: None.

NCEE or REL Study Use:[2] Reading First Impact Study; Improving the Comprehension and Vocabulary Skills of English Language Learners (ELLs) in 5th Grade Using Collaborative Strategic Reading (REL-Southwest)

[1] The reliability rating refers to internal consistency reliability, which was required for direct assessments (see Appendix A). See the reliability section in the profile narrative for information available on test-retest and alternate form reliability.

[2] See Table F.1 for web address.

References:

Guilford, J.P., and Ralph Hoepfner. *The Analysis of Intelligence*. New York: McGraw-Hill, 1971.

Mather, Nancy, Donald D. Hammill, Elizabeth A. Allen, and Rhia Roberts. *Test of Silent Word Reading Fluency. Examiner's Manual*. Austin, TX: PRO-ED, Inc., 2004.

Mee Bell, Sherry, Steve R. McCallum, Bobbie Burton, Rebecca Gray, Sunny Windingstad, and Jessica Moore. "Concurrent Validity of the Test of Silent Word Reading Fluency." *Assessment for Effective Intervention*, vol. 31, no. 3, 2006, pp. 1-9.

Meeker, Mary N., and Robert Meeker. *Structure of Intellect Learning Abilities Test*. El Segundo, CA: SOI (Structure of Intellect) Institute, 1975.

Miller-Guron, Louise. "Wordchains: A Matched English Version of the Swedish Wordchains Test." Unpublished test. University of Goteborg, Sweden: 1996.

Taylor, S.E., H. Frackenpohl, C.E. White, B.W. Nieroda, C.L. Browning, and E.P. Birsner. *EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies*. Orlando, FL: Steck-Vaughn, 1989.

Young, John W. "Review of the Test of Silent Word Reading Fluency." In *The Sixteenth Mental Measurements Yearbook*, edited by Robert A. Spies and Barbara S. Plake. Lincoln, NE: The Buros Institute of Mental Measurements, 2005.

TEST OF WORD READING EFFICIENCY (TOWRE), 1999

Authors: Joseph K. Torgesen, Richard K. Wagner, and Carol A. Rashotte		**Type of Assessment:** Individual assessment **Domain:** Reading (phonemic decoding and sight word vocabulary)
Publisher: PRO-ED, Inc. 800-897-3202 http://www.proedinc.com/		**Grade/Age Range:** 6 through 24 years, 11 months **Administration Interval:** No information except for "regular intervals during 1st and 2nd grade"
Material, Training, and Scoring Costs: TOWRE kit (manual; 25 Profile/Examiner Record Booklets for each Form (A and B); Word Cards for each form; and a storage box): $184		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2 (bachelor's degree with coursework in measurement and domain) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours) The assessor should have thorough knowledge of the manual and standard assessment procedures and should be aware of policies regarding test administration, interpretation, and confidentiality. An experienced assessor should observe three or more practice tests.
Languages: English		**Alternate Forms:** Two forms; administration interval not specified
Representativeness of Norming Sample: The norming sample includes 1,507 individuals age 6 through 24 years (106 to 155 6- to 13-year-olds in 1-year intervals, 77 to 93 14- to 17-year-olds in 1-year intervals, and 112 18- to 24-year-olds). The sample was stratified by age; demographic characteristics (region of country, gender, race/ethnicity, rural/urban residence, family income, parent education, and disability status) of the school-age sample are comparable to the 1997 *Statistical Abstract of the United States.*[1] Testing was conducted across 30 states in fall 1997 and spring 1998.		**Summary** **Initial Material Cost:** 2 ($100 to $200) **Time to Administer:** Approximately 5 minutes **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 1[2] (none described) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 2 (older than 10 years or not nationally representative)

NARRATIVE

Description: The TOWRE is an individually administered, timed assessment that measures phonemic decoding and sight word vocabulary; it is normed for individuals age 6 through 24 years. The assessment comprises 167 items that increase in difficulty as the test progresses and consists of two subtests: Sight Word Efficiency (SWE; 104 items), which measures the number of real printed words identified; and Phonemic Decoding Efficiency (PDE; 63 items), which measures the number of pronounceable non-words identified. There are two forms for each subtest, Form A and B. Before the assessor administers the subtests, students must be able to identify orally at least one word/non-word from the practice list. For each subtest, the student receives a list of words or non-words to read as quickly as possible within 45 seconds. Including instructions and practice items, administration of the two subtests requires about 5 minutes (7 to 8 minutes if both forms are administered).

Other Languages: None.

Uses of Information: The TOWRE measures growth in phonemic decoding and sight word reading skills and may be used for research purposes. The developers note that the TOWRE may supplement other measures in diagnosing specific reading disabilities in older students and adults.

Methods of Scoring: Raw scores are calculated by tallying the number of correct responses read within 45 seconds for each subtest.[3] The scores may be converted to age and grade equivalents,[4] percentiles, and standard scores. In addition, tables in the manual provide a selection of total standard scores (in increments of 5 from 55 to 150) that have been converted to normal curve equivalent (NCE) scores, *T*-scores, *z*-scores, and stanines. An individual highly trained on the TOWRE should score the assessment.

Interpretability: The manual provides thorough instructions for interpreting the results of the TOWRE. A skilled assessor is needed to obtain accurate test results; an individual with clinical skills is required for diagnosis. The Profile section of the Examiner Record Booklet provides a graphic representation of total and subtest standard scores.

Reliability:
(1) Internal consistency reliability: According to the authors, Cronbach's alpha and split-half reliability coefficients are inappropriate measures of internal consistency for speeded tests; as a result, the authors used alternate-form reliability (below) to measure internal consistency (Torgesen et al. 1999).
(2) Test-retest reliability: Correlations between the two administrations (interval two weeks of less) ranged from 0.83 to 0.97 across age groups (6- through 9-year-olds, N = 29; and 10- through 18-year-olds, N = 17) for total and subtest standard scores of both forms.
(3) Alternate form reliability: Forms A and B were administered during one testing session. Correlations of raw scores for the SWE, PDE, and total test ranged from 0.86 to 0.98 for students from the norming sample age 6 through 17 years (across 1-year intervals). Alternate form coefficients were also calculated for select subgroups of the norming sample by gender, race/ethnicity, and disability status, with coefficients ranging from 0.92 to 0.98 for total and subtest scores.

To determine whether the two forms measured the same constructs, the developers tested a two-factor model with each form as a factor and rejected the model for poor model fit (i.e., resulting in a significant Chi-square fit statistic and a comparative fit index [CFI] less than 0.83). The results indicated that the two factors did not represent the data, providing evidence that Form A and B measure similar concepts.
(4) Inter-rater reliability: Two staff from the PRO-ED research department independently scored a set of 30 completed protocols randomly selected from the norming sample. Correlations of standard scores between the two staff members for the two subtests and total, respectively, were each 0.99.

Validity Evidence:
Words on the SWE were selected according to word frequency in printed text at the elementary school level, length and complexity of syllables, and number of syllables. *The Reading Teacher's Book of Lists* was a primary source.

Construct/Concurrent validity: Corrected item-total correlations were used for item selection. Subtest median item discrimination coefficients ranged from 0.42 to 0.75 for the SWE across forms and age groups and from 0.48 to 0.64 for the PDE across forms and age groups. Subtest median item difficulty coefficients ranged from 0.15 (6-year-olds) to 0.97 (17-year-olds) for the SWE across forms and age groups and from 0.07 (6-year-olds) to 0.90 (16- and 17-year-olds) for the PDE across forms and age groups. The item selection process differed for each subtest. Items on the PDE subtest were ordered according to grapheme-phoneme combinations (e.g., two-phoneme vowel-consonant [VC] or three phonemes [CVC]). The difficulty level for each subtest, which increases as testing progresses, was determined by using data from the standardization sample whenever possible, although the order of difficulty on the PDE did not consistently follow the grapheme-phoneme ordering described above because difficulty is also measured by the particular phonemes in words and their frequency in the English language. A confirmatory factor analysis was conducted on both forms of the SWE and PDE with data from the norming sample in which good model fit was found (i.e., resulting in an insignificant Chi-square fit statistic and a CFI value of 1.00). The correlation between the latent constructs of the SWE and PDE for the model was 0.84. A one-factor model was also tested but did not show good fit (i.e., resulting in a significant Chi-square fit statistic and a CFI value less than 0.83).

The manual describes several studies comparing scores between the TOWRE and the Woodcock Reading Mastery Tests-Revised (WRMT-R). Scores between the PDE and Word Attack subtest of the WRMT-R ranged from 0.85 to 0.91, and scores between the SWE and Word Identification subtest of the WRMT-R ranged from 0.86 to 0.94.

In other studies, the TOWRE was correlated with the Passage Comprehension subtest from the WRMT-R and the Gray Oral Reading Tests-Third Edition (GORT-3), which measures reading achievement in terms of reading comprehension and accuracy and rate of word reading in context. Correlations ranged from 0.76 to 0.87 between the WRMT-R Comprehension and the SWE and from 0.66 to 0.69 between the WRMT-R Comprehension and the PDE. Correlations ranged from 0.50 to 0.82 between the GORT-3 subtests and the SWE and from 0.47 to 0.75 between the GORT-3 subtests and the PDE. Students in the studies were in grades 1 through 5, with samples ranging from 125 to 201 for those specified.

The authors held an expectation is that chronological age is directly related to performance on the TOWRE. The overall correlations with age were 0.81 for both Form A and B of the SWE and 0.78 and 0.77 for Forms A and B of the PDE, respectively. In addition to age differentiation, scores on the TOWRE were expected to differ by group (e.g., individuals with learning disabilities were expected to have lower scores). Mean total and subtest standard scores across both forms of the assessments ranged from 93.8 to 96.1 for individuals with speech/language handicaps and from 81.9 to 85.1 for individuals with learning disabilities (compared to 99.3 to 100.4 for the norming sample). The authors consider the scores for students with disabilities to be within the normal range, however.

Predictive validity: No information available.

Bias Analysis: A logistic regression analysis was used to detect differential item functioning (DIF) by gender and race/ethnicity across both forms of the assessment. DIF was detected in 28 of the 334 items (10 items on both Forms A and B of the SWE and 3 and 5 items on Forms A and B of the PDE, respectively). Overall, the developers note only a few cases in which group membership is significant, indicating minimal bias in the subtests. Sample sizes and age ranges were not provided. Mean total and subtest standard scores across both forms of the assessment ranged from 95.2 (black students) to 105.9 (Asian students) across race/ethnicity and from 99.1 (males) to 101.5 (females) for the gender subgroup. The authors find the scores for the race/ethnicity and gender subgroups to be within the normal range.

Training Support: The manual provides detailed information on administration, interpretation, and scoring of the TOWRE. Assessors should consult a colleague or supervisor regarding any information in the manual that they do not understand. In addition, an individual familiar with test administration should observe practice administrations and help with scoring and interpretation. For the PDE subtest, assessors are required to demonstrate knowledge of common pronunciation conventions in English in order to avoid scoring errors. The manual provides a table that lists acceptable pronunciation of non-words for assessors familiar with the International Phonetic Alphabet.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: Both subtests of the TOWRE include a Form A and Form B that may be administered during the same testing session. For diagnostic assessment (of older students), both forms should be administered to increase reliability. For cases involving the monitoring of a response to an intervention, only one form should be used during a single administration; several administrations across a year should alternate forms (that is, testing of grade 1 and 2 students four times in a school year should follow a pattern of Form A-Form B-Form A-Form B).

Previous Version: None.

NCEE or REL Study Use:[5] Evaluation of the Effectiveness of Educational Technology Interventions; Closing the Reading Gap

[1] External reviewers have expressed concern about the representativeness of the norming sample (Tindal 2003; Vacca 2003).

[2] The reliability rating refers to internal consistency reliability, which was required for direct assessments (see Appendix A). See the reliability section in the profile narrative for information available on test-retest and alternate form reliability.

[3] The Profile/Examiner Record Booklet provides the correct pronunciation of non-words by using real-word examples, i.e., the "i" in the non-word "ip" is pronounced like the "i" in the word "tip" and the "a" in the non-word "ga" may be pronounced like the "a" in the word "gap" or the "a" in the word "gate."

[4] The authors recommend caution in using age and grade equivalents; standard scores and percentiles are preferable.

[5] See Table F.1 for web address.

References:

Hagan-Burke, Shanna, Mack D. Burke, and Clay Crowder. "The Convergent Validity of the Dynamic Indicator of Basic Early Literacy Skills and the Test of Word Reading Efficiency for the Beginning of First Grade." *Assessment for Effective Intervention*, vol. 31, no. 4, 2006, pp. 1-15.

Tindal, Gerald. "Review of TOWRE." In *The Fifteenth Mental Measurements Yearbook*, edited by Barbara S. Plake, James C. Impara, and Robert A. Spies. Lincoln, NE: The Buros Institute of Mental Measurements, 2003.

Torgesen, Joseph K., Richard K. Wagner, and Carol A. Rashotte. *TOWRE: Test of Word Reading Efficiency*. Austin, TX: PRO-ED, 1999.

Vacca, John J. "Review of TOWRE." In *The Fifteenth Mental Measurements Yearbook*, edited by Barbara S. Plake, James C. Impara, and Robert A. Spies. Lincoln, NE: The Buros Institute of Mental Measurements, 2003.

WOODCOCK-JOHNSON III NORMATIVE UPDATE (WJ III NU), 2007

Authors: Richard W. Woodcock, Kevin S. McGrew, and Nancy Mather		**Type of Assessment:** Individually administered adaptive assessment **Domain:** Cognitive Abilities Battery (COG): intelligence, general knowledge, memory, approaches to learning/motivation; Achievement Battery (ACH): reading (phonological awareness, letter recognition, reading vocabulary, decoding, comprehension), language arts/language proficiency (receptive language, expressive vocabulary, writing, editing skills), mathematics, science, social studies
Publisher: Riverside Publishing 800-323-9540 http://www.woodcock-johnson.com		**Grade/Age Range:** For 7 COG tests and 12 ACH tests, age 2 years through adult; school-age through adult for remaining tests **Administration Interval:** None described
Material, Training, and Scoring Costs: WJ III NU Complete Battery kit (COG and ACH Form A Standard and Extended Test Books, Manuals, Training Workbooks, Audio Recordings, 25 Test Records and Response Booklets; 5 Brief Intellectual Ability Test Records; WJ III NU Compuscore and Profiles Program; Technical Manual; and Scoring Guides): $1,222.50 Achievement Battery kit: $551.25 Cognitive Abilities Battery kit: $775.00		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2+ (certification beyond bachelor's like a master's) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (> 2 hours)
Languages: English, Spanish		**Alternate Forms:** ACH tests have two equivalent forms; administration interval not described.
Representativeness of Norming Sample: The WJ III NU presents an update of normative data for the WJ-III and Batería III cognitive and achievement batteries based on 2005 U.S. Census estimates and updated norming procedures. The updated norms are based on a stratified, nationally representative sample (N = 8,782). Subjects came from 100 geographically diverse U.S. communities. The developers stratified the sample by region, community size, gender, race, Hispanic/non–Hispanic background, foreign-/native-born, and school type.		**Summary** **Initial Material Cost:** 4 (>$500) **Time to Administer:** 45 to 50 minutes COG; 60 to 70 minutes ACH **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 3 (normed within past 10 years and nationally representative)

NARRATIVE

Description: The Woodcock-Johnson III Normative Update (WJ III NU) comprises updated norms and norming procedures for two co-normed assessment batteries, the Woodcock-Johnson III Tests of Cognitive Abilities (WJ III COG) and the Woodcock-Johnson III Tests of Achievement (WJ III ACH). The updates are incorporated into the WJ III NU computer scoring program and Technical Manual. The test administration materials remain the same as for the previous WJ III versions. The instruments provide a comprehensive set of norm-referenced tests for measuring intellectual abilities and academic achievement in individuals age two years through adulthood. The WJ III COG consists of a standard battery of 10 tests, an extended battery of 10 tests (to provide in-depth assessment of different types of abilities), and 11 supplementary diagnostic tests (to pinpoint further any specific areas of weakness or strength). Standard battery COG tests include (1) Verbal Comprehension, (2) Visual-Auditory Learning, (3) Spatial Relations, (4) Sound Blending, (5) Concept Formation, (6) Visual Matching, (7) Numbers Reversed, (8) Incomplete Words, (9) Auditory Working Memory, and (10) Visual-Auditory Learning-Delayed. Standard and extended battery COG tests may be grouped to yield three overall categories of cluster scores: (1) Cognitive Performance clusters (Verbal Ability, Thinking Ability, and Cognitive Efficiency); (2) Cattell-Horn-Carroll (CHC) Factor Clusters (Comprehension-Knowledge, Long-Term Retrieval, Visual-Spatial Thinking, Auditory Processing, Fluid Reasoning, Processing Speed, and Short-Term Memory); and (3) Clinical Clusters (Phonemic Awareness, Working Memory, Broad Attention, Cognitive Fluency, Executive Processes, Delayed Recall, and Knowledge). The WJ III ACH consists of a standard battery of 12 tests and an extended battery of 10 tests (to provide in-depth assessment of an achievement area). Standard battery ACH tests include (1) Letter-Word Identification, (2) Reading Fluency, (3) Story Recall, (4) Understanding Directions, (5) Calculation, (6) Math Fluency, (7) Spelling, (8) Writing Fluency, (9) Passage Comprehension, (10) Applied Problems, (11) Writing Samples, and (12) Story Recall-Delayed. The WJ III NU also introduced the WJ III Tests of Achievement Form C/Brief Battery (WJ III Form C/Brief Battery) offering abbreviated achievement testing options through selected achievement tests. Across all batteries, the assessor may tailor the administration by selecting the tests that best tap the abilities and skills of interest for a particular student. Each test takes approximately 5 minutes, with the COG standard battery requiring 45 to 50 minutes and the ACH requiring 60 to 70 minutes. Floor and ceiling effects have been observed on some WJ III tests with students age 2 years and 5 years, 6 months and with 16- to 25-year-old students (Bradley-Johnson and Durmusoglu 2005; Krasa 2007).

Other Languages: The Batería III Woodcock-Muñoz is a Spanish adaptation of the WJ III that allows for comprehensive assessment of intellectual ability (including bilingual and low verbal ability), specific cognitive abilities, scholastic aptitude, oral language, and academic achievement in individuals age 2 to 90 years. All tests of the WJ III have been translated or adapted into Spanish for the Batería III. For the Batería III Woodcock-Muñoz: Pruebas de habilidades cognitivas (Batería III COG), assessors may choose from six scales: (1) brief, (2) standard, (3) extended, (4) early development, (5) bilingual (with Diagnostic Supplement), and (6) low verbal (with Diagnostic Supplement). The Batería III Woodcock-Muñoz: Pruebas de aprovechamiento (Batería III APROV) consists of five reading tests, four oral language tests, four mathematics tests, four written language tests, and four supplemental tests of academic language proficiency. The Comparative Language Index (CLI) may also be used to assess language dominance. The WJ III NU computer scoring program provides updated norms for the Batería III; in addition, a

Spanish version of the Woodcock Interpretation and Instructional Interventions Program software is available (see Interpretability). Given that score scales are linked with those of the WJ III, individual scores on the Batería III may be compared directly to WJ III scores. Such comparability is useful for comparing students' proficiency on assessed tasks in both Spanish and English. The computer scoring program may also compute students' cognitive-academic language proficiency (CALP).

The developers collected data from a calibration sample of 1,413 native Spanish speakers from various Spanish-speaking regions (279 were from nine U.S. states). Using Item Response Theory (IRT) methods, developers equated Batería III test data to that of parallel tests on the WJ III, making the scores between the WJ III and the Batería III directly comparable.

Schrank et al. (2005) reported that confirmatory factor analyses (CFA) of the Batería III standardization data supported the measure's CHC theory-based latent factor structure (one general factor and nine broad factors) with subsamples of 6- to 13-year-olds and 14- to 19-year-olds. Patterns and magnitudes of Batería III factor loadings demonstrated a latent factor structure similar to that of the WJ III. Batería III internal consistency reliability coefficients for scores approximated those of the WJ III norming sample. For 4- to 13-year-olds, coefficients ranged from 0.72 to 0.94 on cognitive battery tests and from 0.67 to 0.98 on achievement tests (Schrank et al. 2005).

Uses of Information: The WJ III NU permits age- or grade-based norm-referenced interpretation for individual ability and achievement scores. The information may be used for diagnosis of academic strengths and weaknesses, educational programming, growth assessment, program evaluation, and research.

Methods of Scoring: The Examiner's Manuals and the test easels (the flip books used for testing) summarize the general test and individual item scoring rules. The assessor indicates on the test record form whether the child passes or fails an item. The assessor computes raw scores by summing the number of correct responses and then enters the scores into the computer scoring program, which generates norm-referenced scores (computer scoring is required for the WJ III NU). Grade or age equivalents, instructional ranges, standard scores (deviation quotients), and percentile ranks may be computed for each test and cluster. Users may also compute relative proficiency indexes (RPI), which are ratios reflecting a person's performance compared to the performance of the average student of the same age or grade.

Interpretability: The Examiner's Manuals provide information about how to interpret individual test scores, cluster scores, and discrepancies between scores in the cognitive and ability areas. The WJ III NU computer scoring program offers options for interpreting intra-individual profiles of cognitive abilities and achievement as well as ability-achievement discrepancies. The Woodcock Interpretation and Instructional Interventions Program software provides assistance with test interpretation, linking assessment results to evidence-based interventions and report writing.

Reliability:
(1) Internal consistency reliability: The developers calculated split-half reliability estimates for scores for all tests, except the timed tests and tests with multiple-point scoring systems, for

which they conducted Rasch analysis procedures. For the 31 WJ III NU Tests of Cognitive Abilities, reliability estimates for scores ranged from 0.61 to 0.99 for 5- to 18-year-olds (calculated separately for each year of age), with most estimates at 0.80 or above. For the achievement tests, reliability estimates for scores ranged from 0.57 to 0.99; again, most estimates were at 0.80 or above. The publishers recommend the use of cluster scores (i.e., groups of items from two or more tests) because such scores demonstrate consistently higher reliability.

(2) Test-retest reliability: The WJ III NU Technical Manual reports results of three test-retest reliability estimation studies. First, researchers computed the test-retest reliability estimates of the 9 WJ III NU speeded tests with one-day intervals. The reliability estimates ranged from 0.75 to 0.94 for students age 7 to 11 years and from 0.72 to 0.97 for students age 14 to 17 years. Another study reported test-retest reliability estimates for 15 cognitive and achievement tests, with intervals ranging from less than 1 year to more than 10 years (the authors described the estimates as evidence of extended test-retest reliability, but many researchers would interpret them as evidence of predictive validity). For retest intervals between one and two years, reliability estimates ranged from 0.62 to 0.94 for students age 2 to 18 years (most were above 0.70); for retest intervals from 3 to 10 years with the same age group, estimates ranged from 0.35 to 0.92, with tests in the Thinking Abilities cluster exhibiting lower reliability estimates than those in the Acquired Knowledge cluster. In a third study conducted with 457 students age 4 to 17 years, researchers calculated the test-retest reliability estimates of the 17 WJ III ACH tests and 16 clusters, with a retest interval of one year. Reliability coefficients across ages ranged from 0.69 to 0.96 for scores across the tests and from 0.93 to 0.99 for the clusters.

(3) Alternate form reliability: The WJ III NU Technical Manual cites evidence from the WJ III ACH of alternative form reliability estimates for selected subtests on Forms A, B, and C. For the Calculation subtest, the authors cited similar item difficulties ($r = 0.99$ between Forms A and C), response ogives (item curves by ability for a given item difficulty), and standard errors by level of ability across forms as evidence of reliability and construct validity. For Passage Comprehension, the median alternate form correlation between Forms A and B was 0.85 across all age groups; most correlations ranged from 0.85 to 0.96.

(4) Inter-rater reliability: Three inter-rater reliability studies of subjective ratings of responses on the Writing Samples test were conducted with elementary and high school students and students with learning disabilities for the Woodcock Johnson-Revised (WJ-R) norming sample. The first two studies yielded intercorrelations ranging from 0.89 to 0.98 (to correct for length, median Spearman-Brown correlations ranged from 0.90 to 0.98, respectively). The third study reported an intercorrelation of 0.93 among four raters' ratings of responses for 47 students with learning disabilities (Mather et al. 1991).

Validity Evidence:

The tests and clusters are based on the CHC theory of cognitive abilities. The WJ III NU's content rests on its adherence to CHC theory. Content was also designed to assess core curricular areas and areas specified in federal legislation. For the cognitive battery, experts developed test items to measure both narrow and broad abilities; each test is intended to measure a discrete narrow ability, and clusters of tests are meant to assess broad abilities. Achievement test items were developed to sample skills of oral language and academic achievement in reading, mathematics, written language, and curricular knowledge. The Technical Manual cites data demonstrating the growth and decline of cognitive and achievement abilities across the lifespan.

Construct/Concurrent validity: Developers conducted two confirmatory factor analyses with the WJ III norming sample (N = 3,900). Results indicated that, for the WJ III COG, a latent factor model with one general factor (*g*) and seven broad factors provided the best fit among alternative models. Data analyses on the combined cognitive and achievement batteries showed that an expanded model with one general factor and nine broad factors, plus several narrow abilities, provided the most plausible fit.

The WJ III Technical Manual cites positive correlations between WJ III tests and clusters measuring similar constructs. For example, correlations among the Comprehension-Knowledge tests of Verbal Comprehension, General Information, Picture Vocabulary, and Academic Knowledge ranged from 0.61 to 0.90 for 4- to 19-year-olds. Similar correlations were observed among clusters measuring related abilities. The developers found positive correlations between WJ III tests and clusters and other tests measuring similar constructs. Studies conducted with preschool and elementary-age samples found correlations in the 0.70s between the WJ III General Intellectual Ability standard and extended scores with full-scale or composite scores from several widely used aptitude tests. These studies also reported correlations ranging from 0.60 to 0.70 between the WJ III Brief Intellectual Ability score and other aptitude tests. For students in grades 1 through 8, the WJ III ACH Total Achievement Score correlated with overall achievement scores on the Wechsler Individual Achievement Test (WIAT) and the Kaufman Test of Educational Achievement (KTEA) (r = 0.65 and 0.79, respectively). Higher correlations were observed when comparing scores within specific academic areas across these tests.

The WJ III Technical Manual states that intercorrelations among tests measuring different abilities were lower than those between tests measuring similar abilities. For example, correlations between the Comprehension-Knowledge tests of Verbal Comprehension, General Information, Picture Vocabulary, and Academic Knowledge with the Visual-Spatial tests of Spatial Relations and Picture Recognition ranged from 0.13 to 0.48 for 4- to 19-year-olds. At the cluster level, intercorrelations among WJ III COG clusters typically ranged from 0.20 to 0.60 across age groups.

The Technical Manual presents median scores and standard deviations for selected WJ III NU clusters for the total norming sample and for 11 clinical samples comprising individuals with developmental, educational, and neuropsychological disabilities and gifted students. Although statistical significance was not established, cluster score differences were observed across clinical groups. For example, gifted students' median cluster scores ranged from 103 to 121 versus 99 to 101 for the total norming sample.

Predictive validity: See above discussion of extended test-retest reliability.

Bias Analysis: The developers conducted bias analyses during the development of the WJ III to minimize potential bias related to gender, race, Hispanic origin, and disability status. First, experts reviewed items for potential bias and eliminated or revised all items identified as potentially biased. Next, selected items were subjected to differential item functioning (DIF) analyses conducted with the Rasch Item Response Theory (IRT) model. The analyses focused on a pool of items from the WJ III COG Comprehension-Knowledge tests and the WJ III ACH Academic Knowledge test in view of the tests' strong emphases on language and achievement influences. The items assessed vocabulary, general language development, general information,

and curricular and general cultural knowledge. The results indicated that only a few items differed significantly between groups; expert reviewers flagged and removed one of the items. The other items were retained because of the possibility of spurious findings related to the number of statistical comparisons conducted. In addition, the developers conducted three multiple-group CFAs to examine latent factor structure invariance across groups. The latent factor structure of the WJ III was largely invariant between males and females, White and non-White students, and Hispanic and non-Hispanic student. An independent study supporting the latent structural invariance of the WJ III for Black and White students (Edwards and Oakland 2006) replicated this finding.

Training Support: Training videos and workbooks are available from the publisher. The publisher offers national and regional group training sessions (typically costing from $1,500 to $2,000 for one day of training) as well as individual training sessions. Technical support is available by telephone and online.

Adaptations/Special Instructions for Individuals with Disabilities: The Examiner's Manuals describe accommodations for testing individuals with various difficulties and impairments (including attentional, behavioral, reading, hearing, visual, and physical disabilities).

Alternate Forms: The WJ III ACH includes two forms (A and B) for all 22 oral language and achievement tests. The WJ III NU also introduces the WJ III Form C/Brief Battery consisting of 9 achievement tests in reading, mathematics, and written language (3 of each). Users may now compute Brief Achievement, Brief Reading, Brief Math, and Brief Writing cluster scores for all three forms. The Examiner's Manual does not describe a recommended time interval between administrations of alternate forms.

Previous Version: Major differences between the WJ III NU and the WJ III include updated norms and norming procedures for the WJ III and Batería III cognitive and achievement batteries. For the WJ III ACH, the WJ III Form C/Brief Battery makes Brief Reading, Brief Math, and Brief Writing cluster scores available for Forms A and B. The WJ III NU also includes updated validity information as well as new methods of analyzing intra-individual variation in cognitive and achievement performance. The WJ III NU Compuscore and Profiles Program include new parent report and summary report options that present findings in terms of standard or proficiency scores.

NCEE or REL Study Use:[1] An Impact Evaluation of Early Literacy Programs: The Effects of Opening the World of Learning (OWL) on the Early Literacy Skills of At-Risk Urban Preschool Students (REL-Appalachia); Program for Infant and Toddler Caregivers (PITC) (REL-West)

[1] See Table F.1 for web address.

References:

Bradley-Johnson, Sharon, and Gokce Durmusoglu. "Evaluation of Floors and Item Gradients for Reading and Math Tests for Young Children." *Journal of Psychoeducational Assessment*, vol. 23, no. 3, 2005, pp. 262-278.

Edwards, Oliver, and Thomas Oakland. "Factorial Invariance of Woodcock-Johnson III Scores for African Americans and Caucasian Americans." *Journal of Psychoeducational Assessment*, vol. 24, no. 4, 2006, pp. 358-366.

Kisker, Ellen E., Kimberly Boller, Charles Nagatoshi, Christine Sciarrino, Vinita Jethwani, Teresa Zavitsky, Melissa Ford, and John M. Love. "Resources for Measuring Services and Outcomes in Head Start Programs Serving Infants and Toddlers." Washington, DC: Mathematica Policy Research, 2003.

Krasa, Nancy. "Is the Woodcock-Johnson III a Test for All Seasons? Ceiling and Item Gradient Considerations in Its Use With Older Students." *Journal of Psychoeducational Assessment*, vol. 25, no. 1, 2007, pp. 3-16.

Mather, Nancy, and Richard W. Woodcock. *Examiner's Manual. Woodcock-Johnson III Tests of Cognitive Abilities*. Itasca, IL: Riverside Publishing, 2001.

McGrew, Kevin S., Fredrick A. Schrank, and Richard W. Woodcock. *Technical Manual. Woodcock-Johnson III Normative Update*. Rolling Meadows, IL: Riverside Publishing, 2007.

Muñoz-Sandoval, Ana F., Richard W. Woodcock, Kevin S. McGrew, and Nancy Mather. *Batería III Woodcock-Muñoz*. Itasca, IL: Riverside Publishing, 2005a.

Muñoz-Sandoval, Ana F., Richard W. Woodcock, Kevin S. McGrew, and Nancy Mather. *Batería III Woodcock-Muñoz: Pruebas De Aprovechamiento*. Itasca, IL: Riverside Publishing, 2005b.

Muñoz-Sandoval, Ana F., Richard W. Woodcock, Kevin S. McGrew, and Nancy Mather. *Batería III Woodcock-Muñoz: Pruebas De Habilidades Cognitivas*. Itasca, IL: Riverside Publishing, 2005c.

Schrank, Fredrick A., Kevin S. McGrew, Mary L. Ruef, Criselda G. Alvarado, Ana F. Muñoz-Sandoval, and Richard W. Woodcock. "Overview and Technical Supplement (Batería III Woodcock-Muñoz Assessment Service Bulletin No. 1)." Itasca, IL: Riverside Publishing, 2005.

Schrank, Fredrick A., and Richard W. Woodcock. *WJ III Normative Update Compuscore and Profiles Program (Version 3.0) [Computer Software]*. *Woodcock-Johnson III*. Rolling Meadows, IL: Riverside Publishing, 2007.

Woodcock, Richard W., Kevin S. McGrew, and Nancy Mather. *Woodcock-Johnson III*. Rolling Meadows, IL: Riverside Publishing, 2001.

Woodcock, Richard W., Kevin S. McGrew, and Nancy Mather. *Woodcock-Johnson III Tests of Cognitive Abilities*. Rolling Meadows, IL: Riverside Publishing, 2001, 2007.

Woodcock, Richard W., Kevin S. McGrew, Nancy Mather, and Fredrick A. Schrank. *Woodcock-Johnson III Diagnostic Supplement to the Tests of Cognitive Abilities*. Rolling Meadows, IL: Riverside Publishing, 2003, 2007.

Woodcock, Richard W., Kevin S. McGrew, Fredrick A. Schrank, and Nancy Mather. *Woodcock-Johnson III Normative Update*. Rolling Meadows, IL: Riverside Publishing, 2001, 2007.

Woodcock, Richard W., Fredrick A. Schrank, Nancy Mather, and Kevin S. McGrew. *Woodcock-Johnson III Tests of Achievement, Form C/Brief Battery*. Rolling Meadows, IL: Riverside Publishing, 2007.

WOODCOCK READING MASTERY TESTS-REVISED/NORMATIVE UPDATE (WRMT-R/NU), 1998

Authors: Richard W. Woodcock		**Type of Assessment:** Individually administered adaptive assessment **Domain:** Reading (letter identification, reading vocabulary, comprehension)
Publisher: Pearson Assessments 800-627-7271 http://www.pearsonassessments.com		**Grade/Age Range:** Kindergarten through grade 16 or age 5 through 75+ years **Administration Interval:** Frequent with alternate forms
Material, Training, and Scoring Costs: WRMT-R/NU kit (test books/easel, 25 test records, sample NU form, summary record form, Pronunciation Guide Cassette, Sample Report to Parents, NU Examiner Manual, and carry bag): $334.75 for each form; $489.25 for both Forms G[1] and H WRMT-R/NU ASSIST scoring: $249, or $452.25 with one form kit, $606.75 with kit containing both forms		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2 (bachelor's degree with coursework in measurement and domain) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours)
Languages: English		**Alternate Forms:** Two;[1] "frequent re-testing"
Representativeness of Norming Sample: The WRMT-R normative update assessed 3,184 students[2] in kindergarten through grade 12 (approximately 200 to 300 students in each grade) stratified for a nationally representative sample to match the U.S. Census's 1994 *Current Population Survey* based on grade, gender, socioeconomic status, parent education, race/ethnicity, and region. Students with gifted or special education status were included. Students not proficient in English were not included. Students were located at 129 sites in 40 states during 1995 and 1996.		**Summary** **Initial Material Cost:** 3 ($200 to $500) **Time to Administer:** 30 to 45 minutes **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 2 (older than 10 years or not nationally representative)

NARRATIVE

Description: The WRMT-R/NU is an individually administered, adaptive assessment of reading ability that measures decoding and comprehension as well as the skills deemed necessary for beginning reading (such as letter identification). The measure is designed for students in kindergarten through college as well as for adults through age 75+ years. It requires about 30 to 45 minutes for administration, with about 5 to 10 minutes for each of six subtests (referred to by the author as tests), using easel administration with students who give oral responses. The WRMT-R/NU comprises two main areas—reading achievement and readiness. Reading achievement (referred to as the Total Reading Full Scale) contains two clusters, each made up of two subtests. The Basic Skills cluster includes the Word Identification (106 items, isolated words) and Word Attack (45 items, nonsense words) subtests. The Reading Comprehension cluster involves the Passage Comprehension and Word Comprehension subtests, the latter of which involves three components on Antonyms (34 items), Synonyms (33 items), and Analogies (79 items) whose items also provide measures of subject-specific vocabulary for General Reading, Science-Mathematics, Social Studies, and Humanities. A Total Reading Short Scale may be administered with just the Word Identification and Passage Comprehension subtests. The Readiness cluster, found only in Form G, contains two subtests—Visual-Auditory Learning (reproduced from the Woodcock-Johnson Psycho-Educational Battery) and Letter Identification (51 items in "forms that [the student] has never seen" such as roman, italic, bold, serif, and cursive, p. 5). A supplementary checklist for identifying upper- and lower-case letters is also available for the Readiness cluster. For the subtests, assessors follow suggested starting points based on the student's estimated reading level, which may not necessarily match the grade in which the student is enrolled. All items on an entire easel page are administered. From that starting point, the WRMT-R/NU is then administered until establishing a basal level of six consecutive correct items (the lowest six correct items on a page) and a ceiling level of six consecutive incorrect items (the highest six incorrect items on a page).

Other Languages: None.

Uses of Information: The WRMT-R/NU provides a measure of global reading ability in terms of both status and growth. The author purports that the measure may also be used for evaluating program effectiveness as well as for clinical and research purposes. For example, the author notes that the WRMT-R could provide a clinical assessment to assist in diagnosing reading problems, to develop instructional plans, and to determine placement in instructional groups.

Methods of Scoring: During administration, the assessor scores each item as correct (1) or incorrect (0), along with the exact response for use in later error analysis. Scores may then be calculated for subtests and, in turn, for clusters and total reading scale scores by using both grade and age norms. For most subtests, raw scores reflect the sum of correct answers and include all items not administered below the basal level. Exceptions include the Visual-Auditory Learning and Word Comprehension subtests. For the Visual-Auditory Learning subtest, the raw score is determined by subtracting the total number of errors from a value of 134. For Word Comprehension, which involves three components, raw scores are first calculated separately for each component, converted to a part score based on tables provided, and then summed. Additional information is available in the manual for calculating raw scores for the separate subject vocabularies. Word Attack may be scored for total response or for components of words.

From the raw scores, tables are then available to obtain age equivalent scores (age 5 through 75+ years) and grade equivalent scores (kindergarten through college), standard errors of measurement (SEM), W scores (Rasch-based ability score, 500 = mean item difficulty which for the assessment is approximately grade 5 ability), and reference scores (median W score for a particular grade minus 100 to reduce instances of negative difference scores); the latter two are used to calculate a W-difference score (W score minus reference score). Cluster W scores are the average of the W scores for the appropriate subtests. The W-difference score then is used to obtain age- and grade-based percentile ranks, standard scores (mean = 100, standard deviation = 15), and a Relative Performance Index (mastery level in comparison to 90 percent for a particular grade). Percentile ranks are estimated for the 10th, 50th, and 90th points but extrapolated for other intervals (Crocker 2001). Additional standard score options include normal curve equivalents, *T*-scores, and stanines, which are calculated by using the percentile rank and a conversion table. If both forms are administered at once, a Form G+H test record is used to combine results to "maximize measurement precision and the amount of data available for error analysis" (Woodcock 1998, p. 3). ASSIST scoring software is available to input raw scores and obtain all derived scores.

External reviewers express concerns about the scaling and equating of the new norms (Crocker 2001; Murray-Ward 2001). The use of several assessments in norming limits the use of WRMT-R/NU scores to represent a general performance measure rather than particular aspects of reading or language. Further, given gaps in scores, the Word Identification and Letter Identification subtests were normed without the kindergarten and grade 1 sample, but the manual presents kindergarten/grade 1 norms with other grade levels in the tables (Murray-Ward 2001). The developer notes that the WRMT-R/NU, as compared to previous norms, results in higher standard scores for below-average students. These higher scores may have implications for the potential use of the updated normed scores for clinical decisions (Murray-Ward 2001).

Interpretability: The assessor should be familiar with individual testing and measurement. The Examiner's Manual details appropriate training for use of the WRMT-R results, including chapters on interpretation, implications, and technical properties of the assessment; practice exercises; and a self-evaluation checklist. In addition, the Examiner's Manual describes each type of score, the various profiles, and discrepancy and error analyses and includes a chapter on the steps involved in determining instructional implications. The test record contains profiles for obtaining instructional levels and percentile ranks as well as diagnostic profiles for the Readiness, Basic Skills, and Reading Comprehension clusters.

Reliability:
Information provided on reliability is based on the WRMT-R's standardization sample (4,201 students in kindergarten through grade 12) conducted between 1983 and 1985.
(1) Internal consistency reliability: Reliability coefficients (r_{11}), using a split-half approach with raw scores and corrected for length with a Spearman-Brown formula, were presented for grades 1, 3, 5, 8, and 11. Split-half reliability ranged from 0.92 to 0.99 for the Total Reading Full Scale score across Forms G and H and from 0.86 to 0.99 for the Total Reading Short Scale score. The Basic Skills cluster reliabilities ranged from 0.91 to 0.99 across forms and grades while the Reading Comprehension cluster reliabilities ranged from 0.87 to 0.98. Subtests for the two clusters had split-half reliabilities ranging from 0.84 to 0.99. The Readiness cluster reliabilities differed by grade level—0.96 for grade 1 students, 0.88 for grade 3 students, and 0.54 for grade 5

students. The Readiness cluster subtests ranged in reliability from 0.84 to 0.95, except for Letter Identification among grade 5 students (with a split-half reliability of 0.34).

(2) Test-retest reliability: No information available.
(3) Alternate form reliability: No information available.
(4) Inter-rater reliability: No information available.

Validity Evidence:
Development of the WRMT-R involved outside experts, including teachers and curriculum specialists. The author directs the assessor to the scope and sequence of items in the WRMT-R to support content coverage as well as the open-ended format of items that mimic the task of reading. The author used classical item and Rasch modeling techniques to select items based on statistical criteria. No other information on item sources, development, or statistics was provided.

Construct/Concurrent validity: The WRMT-R's (without the normative update scores) reading achievement subtests were correlated with the Woodcock-Johnson (WJ) reading tests for Letter-Word Identification, Word Attack, and Passage Comprehension. Reported correlations between total scores ranged from 0.85 to 0.91 for samples of students in grades 1, 3, 5, and 8 (sample sizes ranging from 33 to 122). Correlations between Word Identification and the WJ Letter-Word Identification ranged from 0.69 to 0.83 and from 0.48 to 0.76 with the other WJ reading subtests across grades. Correlations between the Word Attack subtests for each test ranged from 0.64 to 0.90 and from 0.35 to 0.71 for the other WJ subtests. Correlations between the WRMT-R and WJ Passage Comprehension subtests ranged from 0.55 to 0.71, and the WRMT-R Passage Comprehension was correlated from 0.25 to 0.66 with the other WJ subtests across grades.

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: The assessor is instructed to study the Examiner's Manual on administration and scoring before administering the assessment. The Examiner's Manual includes practice exercises and a self-evaluation checklist. In addition, the WRMT-R/NU kit includes a Pronunciation Guide Cassette to help the assessor develop his/her skills in discriminating between sounds in order to score items in the Word Attack and Word Identification subtests. The assessor should administer and score at least two practice tests and, if newly trained, should administer the full WRMT-R during observation by an experienced WRMT-R assessor.

Adaptations/Special Instructions for Individuals with Disabilities: Additional or unlimited time is not permitted, but more time for a specific item may be allowed if a student requests it. No other adaptations are noted.

Alternate Forms: The WRMT-R/NU includes two forms—G and H. Form G contains two subtests not found in Form H (Visual-Auditory Learning and Letter Identification). With the two forms, the author notes that frequent retesting is possible. For greater precision, the author suggests administering both forms, using the Form G+H test record.

Previous Version: The WRMT-R/NU is the normative update to the WRMT-R (1987). No changes were made to the easels or items, but test records were revised to include Instructional Level Profiles, Part Score Tables for Word Comprehension subtests, and Diagnostic Profiles. The WRMT-R was a revised edition of the 1973 WRMT; it added the Readiness cluster, Antonyms and Synonyms (to the Word Comprehension subtest), subject area vocabulary analysis, more sample items for practice, the Short Scale option, expanded test records to permit error analysis, and extended norms to include college students and older adults.

NCEE or REL Study Use:[3] Closing the Reading Gap

[1] Form G includes two subtests (Visual-Auditory Learning and Letter Identification) not included in Form H.

[2] The normative update involved co-norming the WRMT-R along with four other measures (e.g., Peabody Individual Achievement Test-Revised [PIAT-R] or the Kaufman Test of Education Achievement [K-TEA]). Not all students completed all measures; rather, each student took one complete test battery with at least one subtest from another battery, but norms are based on the entire sample of students completing any of the measures within a certain domain (e.g., word reading). Norming samples then varied by subtest (to include up to 245 adults age 18 to 22 years): 2,151 for Passage Comprehension, 2,662 for Letter Identification and Word Identification, 1,309 for Visual-Auditory Learning, 751 for Word Attack, and 721 for Word Comprehension. Approximately 675 students from kindergarten through grade 12 took the WRMT-R as a primary battery, with 150 to 200 students each in three-grade intervals (e.g., kindergarten through grade 2). In addition, the norming sample for the update was noted as limited in terms of the number of students from metropolitan areas and did not include adults older than 22 years or college students (Murray-Ward 2001). Thus, the manual includes both old and new norms.

[3] See Table F.1 for web address.

References:

Crocker, Linda. "Review of the Woodcock Reading Mastery Tests-Revised (1998 Normative Update)." In *The Fourteenth Mental Measurements Yearbook*, edited by Barbara S. Plake and James C. Impara. Lincoln, NE: The Buros Institute of Mental Measurements, 2001.

Murray-Ward, Mildred. "Review of the Woodcock Reading Mastery Tests-Revised (1998 Normative Update)." In *The Fourteenth Mental Measurements Yearbook*, edited by Barbara S. Plake and James C. Impara. Lincoln, NE: The Buros Institute of Mental Measurements, 2001.

Pearson Assessments. "WRMT-R/NU Technical Information." Available at [http://www.pearsonassessments.com/PDF/pubsum/WRMT.pdf]. 1998.

Woodcock, Richard W. *Woodcock Reading Mastery Tests-Revised. Normative Update. Forms G and H. Examiner's Manual*. Minneapolis, MN: Pearson Assessments, 1998.

TABLE B.1

NCEE OR REL RECENTLY DEVELOPED STUDENT ACHIEVEMENT/DEVELOPMENT MEASURES SUMMARY

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range [b]	Reliability	Validity
Vocabulary, Communication					
Lexical diversity	Program to Accelerate Vocabulary Development (REL-Southeast) (3)	This measure is an individual assessment of language arts/language proficiency, specifically lexical diversity, which is the number of unique words relative to the total number of words spoken. The assessor guides each student through a 10-minute task whereby the student tells a story from a wordless picture book. The assessor tape records the task. Using the Computerized Language Analysis program, the tape is transcribed and analyzed for the student's lexical diversity. The expected sample size is approximately 640 students from 160 classrooms in 60 to 80 schools.	Kindergarten and grade 1	Not available	The OMB documents note that the measure has been used in previous studies. No other information was provided.
Approaches toward Learning, Motivation					
Student questionnaire of economic interest and attitudes	UCLA/CRESST Problem-Based Economics (REL-West) (3)	This measure is a student self-report of approaches to learning/motivation. Specifically, it measures student interest in learning economics (e.g. interest in reading about economic issues) on a scale from 1 (very interested) to 5 (not interested). In addition to the outcomes noted above, the questionnaire captures students' attitudes toward school, their school behaviors, and their problem-solving skills. The questionnaire also asks students about classroom practices (see Table D.1). The study enrolled 4,800 students from 40 schools.	Grade 12	An earlier brief version of the current questionnaire had a Cronbach's alpha of 0.80. No other information available	The questionnaire uses several items from the Student Assessment of Learning Gains, developed by the Wisconsin Center for Educational Research. In addition, economists outside of the study team reviewed the questionnaires during development.

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.
[c]AddHealth = National Longitudinal Study of Adolescent Health; 21stCCLC = Evaluation of 21st Century Community Learning Centers Program.

TABLE B.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range [b]	Reliability	Validity
Student Time-on-Task and Engagement with Print (STEP)	Reading First (1)	This measure is a classroom observation of approaches to learning/motivation. It measures student engagement during reading instruction. Each STEP observation consists of three six-minute sweeps (with six minute intervals between sweeps). Observers classify every student in the classroom as either on-task or off-task. Then, if a student is on-task and engaged with print, the observer classifies the student as reading connected text, reading isolated text, or writing. The study created a class-level analytic variable on the percentage of students in classrooms during the scheduled reading block who are on-task and/or interacting with print. STEP observations were completed in 248 schools; 1,361 grade 1 and 2 classrooms in 2005 and 1,354 classrooms in 2006.	Grades 1–3	Inter-rater reliability: observers achieved an average 89 percent agreement across all codes using a reliability test tape. Inter-rater reliability for the on-task and engaged-with-print codes was tested using a reliability tape coded by a gold standard rater. The reliability was 92 percent average agreement for Reading Connected Text, 77 percent for Reading Isolated Text, and 96 percent for Writing	Not available

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.
[c]AddHealth = National Longitudinal Study of Adolescent Health; 21stCCLC = Evaluation of 21st Century Community Learning Centers Program.

TABLE B.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range[b]	Reliability	Validity
Social-Emotional Well-Being					
Student questionnaire of behaviors and violence	School-Based Violence Prevention (3, 4)	This measure is a student self-report of socio-emotional behaviors, including aggression, prosocial behaviors, and victimization. Students report on their own behaviors and other students' behaviors toward them. The questionnaire consists of 85 items that include background information, questions about interpersonal relationships at school, and questions about the student's feelings and attitudes. In particular, the outcomes of the questionnaire focus on students' prosocial behaviors (20 items, e.g., "Share something with a kid from your school"), aggression and attitudes about violence (28 items, e.g., "It's OK to hit someone who hit you first"), and victimization by violence or bullying (26 items) as well as on opinions about safety at school (see Table D.1). The questionnaire is a self-administered paper questionnaire, and the responses are almost all yes/no, Likert scale, or checklist responses. The full questionnaire takes approximately 45 minutes for completion. The study includes student questionnaires from approximately 36,920 students in 40 middle schools.	Grades 6–8	Cronbach's alpha for each subscale from the main study data: Aggression: 0.84 (overall), 0.72 (weapons), 0.90 (not weapons) Prosocial behaviors: 0.90 (extended to others), 0.86 (received from others), NA for active intervention subscale Student victimization: 0.89 (overall), 0.84 (overt), 0.84 (relational)	The student questionnaire was pilot tested with fewer than nine students in grades 6, 7, and 8. Except for prosocial behaviors-active intervention, which was developed for this study, the questionnaire uses items adapted from existing instruments, including the Problem Behavior Frequency Scales (Farrell, Kung, White, and Valois 2000); the School Crime Supplement to the National Crime Victimization Survey (U.S. Department of Justice); and the Positive Behavior Scale (Orpinas 2005).

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.
[c]AddHealth = National Longitudinal Study of Adolescent Health; 21stCCLC = Evaluation of 21st Century Community Learning Centers Program.

TABLE B.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range[b]	Reliability	Validity
Social Studies					
Student performance assessment tasks (UCLA\CRESST)	UCLA/CRESST; Problem-Based Economics (REL-West) (3)	This measure is an individual assessment of social studies. Students complete two of five tasks (each requires 20 minutes for completion), assessing their economic conceptual knowledge and economic problem-solving skills. A balanced incomplete block matrix sampling design created five versions of the test booklet. The five tasks are aligned with the units of the Problem-Based Economics (PBE) curriculum used in the study. The scoring rubric assesses each task on the student's quality of understanding, argumentation, misconceptions or errors, and use of prior knowledge. Overall, the study plans to include approximately 4,800 students from 40 schools, with each task completed by approximately 1,920 students.	Grade 12	No information available	The National Center for Research on Evaluation, Standards, and Student Testing at the University of California, Los Angeles developed the tasks and a rubric for scoring the tasks based on experience in similar experimental research. The five tasks were piloted with 300 students.
Other/Multidomain					
Character traits and behavior questionnaire	Lessons in Character Education (REL-West) (3)	This measure is a student self-report of socio-emotional behaviors, measuring attitudes and values consistent with the goals of character education. Adapted from already validated instruments, the questionnaire consists of 81 items and measures student empathy, altruism, school engagement (two scales noted), aggression, delinquent behavior, feeling of belonging, autonomy and influence, and competence. Administration time is about 35 minutes. Approximately 15,000 students in grades 2 through 5 at 50 schools are enrolled in the study, but only those students in grades 4 and 5 will complete the questionnaire.	Grades 4–5	Cronbach's alphas were provided for the subscales based on documentation of the original source instruments. Student empathy: 0.72 Altruism: 0.86 School engagement: 0.75 and 0.86 Aggression: 0.88 Delinquent behavior: 0.71 Feeling of belonging: 0.87 Autonomy and influence: 0.79 Competence: 0.76	The questionnaire uses items and subscales from already validated instruments. The instruments include Funk et al. 2003 (student empathy); Characterplus 2002 (altruism, autonomy and influence, competence); Furrer and Skinner 2003 (school engagement); Opinas and Frankowski 2001 (aggression); and Kisker et al. 2003 (delinquent behavior).

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.
[c]AddHealth = National Longitudinal Study of Adolescent Health; 21stCCLC = Evaluation of 21st Century Community Learning Centers Program.

TABLE B.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range [b]	Reliability	Validity
Student questionnaire of reading behavior and attitudes	Enhanced Reading Opportunities (1)	This measure is a student self-report of reading behaviors and attitudes toward reading. The questionnaire measures students' participation in literacy support activities (e.g., tutoring services with the goal of improving reading and writing skills) and their attitudes toward reading activities. The three main scales derived from the questionnaire on reading behaviors are amount of school-related reading (seven items), amount of non–school-related reading (seven items), and use of reflective reading strategies (four items). The questions on school-related and non–school-related reading ask students to report the number of times they have read different types of text in the past month (from never to every day). The questions on reflective reading strategies ask students to rate their use of strategies (e.g., "I ask myself questions to make sure I know the material that I have been studying for class") from "strongly agree" to "strongly disagree." Other outcomes derived from the questionnaire varied across the study's two cohorts: reading to learn in both cohorts (three to four items); in the first cohort only—positive literacy activities (four items), ease of reading (seven items), persistence on school work (eight items), negative school behavior (four items), and educational aspirations (one item); and in the second cohort only—reading to enjoy (two items) Approximately 2,413 students in the first cohort and 2,171 students in the second cohort from a total of 34 schools completed the questionnaire.	Grade 9	Cronbach's alpha for cohorts 1 (C1) and 2 (C2): Frequency of in-school reading: 0.83 (C1), 0.71 (C2) Frequency of out-of-school reading: 0.73 (C1), 0.75 (C2) Use of reflective reading strategies: 0.88 (C1), 0.77 (C2) Positive literacy activities: 0.76 (C1) Reading to learn: 0.74 (C1), 0.80 (C2) Reading to enjoy: 0.82 (C2) Ease of reading: 0.83 (C1) Persistence on school work: 0.87 (C1) Negative school behavior: 0.71 (C1)	Not available

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.
[c]AddHealth = National Longitudinal Study of Adolescent Health; 21stCCLC = Evaluation of 21st Century Community Learning Centers Program.

TABLE B.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range [b]	Reliability	Validity
Student questionnaire of substance use	Mandatory Random Student Drug Testing (3)	This measure is a student self-report on school engagement, substance use, and attitudes about substance use. The questionnaire asks students about their participation in school activities, their attitudes toward their school, their past drug use, and the likelihood of future drug use. The questions about drug use include frequency scales (from never using a drug to over 40 times), a 5-point scale on likelihood of using drugs in the future (definitely not to definitely will), and a 5-point scale for agreement on attitudes about using drugs (strongly disagree to strongly agree). The questionnaire takes approximately 30 minutes for completion with paper and pen. Approximately 11,400 students in 45 schools will complete the questionnaire.	Grades 9–12	Although the questionnaire uses questions derived from other questionnaires on substance use, information about those items was not available.	Questions were derived from other questionnaires on substance use, including Monitoring the Future. No other information was available.
Student questionnaire on behavior and school	Student Mentoring Program (4)	This measure is a student self-report of socio-emotional behaviors and approaches to learning/motivation. It contains approximately 62 items in nine areas (peer relationships; relationship with parents; relationship with other adults; school bonding/scholastic efficacy; misconduct/delinquent behavior; gang membership; cigarettes, alcohol, and other drugs; personal responsibility/volunteerism; and future orientation). Five subscales are reported: (1) home and community involvement; (2) future orientation; (3) misconduct; (4) delinquency; and (5) scholastic efficacy and school bonding. Students responded on a 1–4 Likert scale for agreement for all items except for 1 dichotomous item on gang involvement (yes/no) and 4 items on delinquency that collect six categories of frequency of use of drugs, alcohol, and tobacco from "never used" to "10 or more tifomes." The study includes approximately 2,400 students in 32 organizations.	Grades 4–8	Cronbach's alpha by subscale for the current study: Home and community involvement: 0.69 Future orientation: 0.76 Misconduct: 0.72 Delinquency: 0.74 Scholastic efficacy and school bonding: 0.72	Some of the items in the questionnaire are adapted from existing sources while some are original items. The sources used by the study include the Self Perception Profile for Adolescents (Harter 1988)*;* Seattle Social Development Project (Hawkins et al. 2001); mentoring research (Rhodes 2003); and items from existing large-scale studies and questionnaires (AddHealth, 21stCCLC, Monitoring the Future, and the Michigan State University Early Adolescent Sur vey II). [c]

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.
[c]AddHealth = National Longitudinal Study of Adolescent Health; 21stCCLC = Evaluation of 21st Century Community Learning Centers Program.

RECENTLY DEVELOPED STUDENT MEASURES REPORT REFERENCES

Corrin, William, Marie-Andrée Somers, James J. Kemple, Elizabeth Nelson, Susan Sepanik, Terry Salinger, and Courtney Tanenbaum. "The Enhanced Reading Opportunities Study: Findings from the Second Year of Implementation." (NCEE 2009-4036). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, November 2008.

Gamse, Beth, Howard Bloom, James Kemple, Robin T. Jacob, Beth Boulay, Laurie Bozzi, Linda Caswell, Megan Horst, W. C. Smith, Robert St. Pierre, Fatih Unlu, Corinne Herlihy, and Pei Zhu. "Reading First Impact Study: Interim Report." (NCEE 2008-4016). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, April 2008.

Gamse, Beth C., Robin T. Jacob, Megan Horst, Beth Boulay, and Fatih Unlu. "Reading First Impact Study Final Report." (NCEE 2009-4038). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, November 2008.

Kemple, James J., William Corrin, Elizabeth Nelson, Terry Salinger, Suzannah Herrmann, and Kathryn Drummond. "The Enhanced Reading Opportunities Study: Early Impact and Implementation Findings." (NCEE 2008-4015). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, January 2008.

APPENDIX C

TEACHER KNOWLEDGE MEASURE PROFILES AND TABLE OF RECENTLY DEVELOPED MEASURES

ASSESSING TEACHER LEARNING ABOUT SCIENCE TEACHING (ATLAST)TEST OF FORCE AND MOTION, 2008

The Assessing Teacher Learning About Science Teaching (ATLAST) Test of Force and Motion has a student version and thus is included in Appendix B, Student Achievement/Development Measures. Please refer to Appendix B for this profile.

DIAGNOSTIC CLASSROOM OBSERVATION TOOL (DCO), 2008

Authors: Nicole Saginor		**Type of Assessment:** Classroom observation **Domain:** Classroom quality (teacher-student interactions and classroom environment), instructional practices (literacy and mathematics/science), and pedagogical content knowledge (literacy and mathematics/science)
Publisher: Corwin Press 800-233-9936 http://www.corwin.com		**Grade/Age Range:** Kindergarten through grade 12 **Administration Interval:** As frequently as desired
Material, Training, and Scoring Costs: *Diagnostic Classroom Observation: Moving Beyond Best Practice*: $31.95 (paperback) Training costs vary.		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2+ (certification beyond bachelor's like a master's) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours) The observer is advised to become thoroughly familiar with the elements of the DCO and to conduct practice observations with colleagues to establish inter-rater agreement.
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (under $100) **Time to Administer:** 45 to 90 minutes per observation **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 [1, 2] (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available[2] **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Diagnostic Classroom Observation (DCO) is designed to assess the literacy and mathematics/science instruction of kindergarten through grade 12 teachers. The two primary versions of the DCO are (1) the Literacy Version (17 items) and (2) the Math/Science Version (28 items). The Literacy Version assesses teachers' efforts to promote literacy development; it is designed for use with kindergarten through grade 5 teachers and upper-level English teachers. The Math/Science Version evaluates teachers' scientific instructional practices and may be used with all grade levels. A Composite Version of the DCO is also available. It contains four additional items related to the use of literacy as a tool to enhance learning in middle school and high school classrooms and may be used to assess any content area. The Composite Version is constructed by adding these four items to the Math/Science Version. For all versions, several observations of 45 to 90 minutes are advised (N. Saginor, personal communication, January 13, 2009). Each version assesses the extent of evidence (none to extensive) observed along three dimensions of instruction: (1) Implementation of the lesson, (2) Content of the lesson, and (3) Classroom Culture. The Implementation section focuses on the effectiveness of instruction and student engagement that occurs during the lesson. Both teacher and student activity are observed. The Content section addresses the teacher's accuracy, level of abstraction, connections to other concepts, and ability to correct student misconceptions. Finally, the Classroom Culture section assesses the learning environment, the level of student engagement, the nature of the working relationships between the teacher and student and among the students themselves, and issues of student equity (equal teacher attention and provision of student supports as well as participation in group work/activities). In addition to the three dimensions of instruction mentioned above, the DCO provides guidance on assessing a teacher's planning and organization of a lesson before the observation.

Other Languages: None.

Uses of Information: The DCO was designed primarily as a tool for school principals to evaluate teachers' lessons in order to facilitate improvement in instructional practices. The DCO may also be used for research purposes to assess the effectiveness of an intervention. In addition, the author notes that the DCO may be used to facilitate effective hiring practices and as a tool to support teacher preparation initiatives.

Methods of Scoring: Each item represents an indicator of quality practice, and *Diagnostic Classroom Observation—Moving Beyond Best Practice* (Saginor 2008) provides specific examples of what to look for during the observation. In addition, the book contains detailed explanations and vignettes illustrating the items. For each item, the extent of evidence for the particular indicator is rated on a five-point scale (no evidence, little evidence, moderate evidence, consistent evidence, extensive evidence). The book also provides general descriptions of each rating along with scenarios depicting each score. On the observation form, the observer provides examples to justify his/her ratings.

Interpretability: The DCO author's book describes how principals may interpret the data for use in post-observation conferences but provides no information on interpreting the data for research purposes.

Reliability:[2]
(1) Internal consistency reliability: The Literacy and Math/Science versions were adapted for use in a teacher preparation study conducted by Mathematica Policy Research (Mathematica) in 2004 (Constantine et al. 2009). Mathematica assessed the internal consistency of scores, obtaining a reliability estimate ranging from 0.88 to 0.98. MPR also assessed internal consistency reliability with data from a 2006 teacher induction study for the Literacy Version (Glazerman et al. 2008). Reliability estimates for scores were 0.89 for Implementation, 0.80 for Content, and 0.93 for Classroom Culture.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: Not applicable.
(4) Inter-rater reliability: No information available. In the 2006 teacher induction study conducted by Mathematica (Glazerman et al. 2008), to achieve certification, observers had to come within 0.75 points of the "gold standard" average rating for the three constructs (Implementation, Content, and Classroom Culture). Inter-rater reliability was not assessed with study teachers after the certification process.

Validity Evidence:[2]
The DCO is based on a review of the research related to instructional practices and three existing teacher assessments (see Saginor 2008 appendix for cross-walk).

Construct/Concurrent validity: A factor analysis conducted using data from Mathematica's teacher induction study (Glazerman et al. 2008) confirmed the measure's theoretical groupings (Implementation, Content, and Classroom Culture), although a higher-order factor may be involved.

Mathematica's teacher induction study researchers found that scores on the Literacy Version of the measure were significantly positively associated with gains in students' outcomes. For every one-point increase in teachers' scores on the original version of the DCO (the Vermont Classroom Observation Tool [VCOT]), students' test score gains increased four to six points (Glazerman et al. 2008).

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: Given that the DCO covers a wide grade range and varied content areas, the author advises training to adapt the instrument for the specific class to be observed (N. Saginor, personal communication, January 13, 2009). The author also recommends the establishment of inter-rater reliability before conducting classroom observations. The author can provide additional training on the instrument to help establish a coding standard. The cost of the training varies with the intended use of the DCO, the number of trainees, and the level of desired reliability (N. Saginor, personal communication, September 29, 2008). In the 2006 teacher induction study conducted by Mathematica (Glazerman et al. 2008), observers had to become certified before they could conduct observations for the study. The DCO's author provided nine days of training for the prospective observers. The observers practiced coding by using videotaped classes and conducted practice observations in the field. Inter-rater reliability was assessed by comparing observers' ratings from a videotaped class to those of the "gold standard"

panel. To achieve certification, observers had to come within 0.75 points of the "gold standard" average rating for the three constructs (Implementation, Content, and Classroom Culture).

Adaptations/Special Instructions for Individuals with Disabilities: Not applicable.

Alternate Forms: Not applicable.

Previous Version: The original version of the DCO was the VCOT developed by the Vermont Institutes (formerly the Vermont Institute for Science, Math and Technology), a nonprofit collaboration among government, business, and education leaders. The VCOT, based on work by the Science and Math Program Improvement project of Western Michigan University and Horizon Research, Inc., originally assessed teachers' instructional practices in mathematics and science and their use of technology. The tool was further revised and piloted with the assistance of the Education Development Center, an international education nonprofit, and the Northeast and Islands Resources for Technology in Education Consortium, one of 10 regional technology-in-education consortia funded by the U.S. Department of Education. As the institute's work expanded into literacy efforts, they developed the Literacy Version of the VCOT. The Composite Version was created for the CRISS project. The teacher induction and teacher preparation studies mentioned above used an adapted version of the VCOT. After the Vermont Institutes reorganized in 2007, the author continued to refine the tool and published it as the DCO.

NCEE or REL Study Use:[3] Impacts of Comprehensive Teacher Induction; An Evaluation of Teachers Trained Through Different Routes to Certification

[1] Only available for the average score of the three components.

[2] The reliability and validity of the DCO scores were assessed in two studies which used a previous version of the measure, the Vermont Classroom Observation Tool (VCOT). Please see Previous Version for information on the VCOT. The DCO's reported psychometric properties are based on observations of elementary school teachers only.

[3] See Table F.1 for web address.

References:

Constantine, Jill, Daniel Player, Tim Silva, Kristin Hallgren, Mary Grider, and John Deke. "An Evaluation of Teachers Trained through Different Routes to Certification, Final Report." (NCEE 2009-4043). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

Glazerman, Steven, Sarah Dolfin, Martha Bleeker, Amy Johnson, Eric Isenberg, Julieta Lugo-Gil, Mary Grider, and Edward Britton. "Impacts of Comprehensive Teacher Induction: Results from the First Year of a Randomized Controlled Study." (NCEE 2009-4034). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, October 2008.

Saginor, Nicole. *Diagnostic Classroom Observation: Moving Beyond Best Practice*. Thousand Oaks, CA: Corwin Press, 2008.

Saginor, Nicole, and Phil Hyjek. *Observing Standards-Based Classrooms: The Vermont Classroom Observation Tool (VCOT)*. Montpelier, VT: Vermont Institutes, March 2005.

PEDAGOGICAL CONTENT KNOWLEDGE ASSESSMENT (PCK), 2008

Authors: Heather C. Hill, Deborah Loewenberg Ball, and Stephen G. Schilling		**Type of Assessment:** Group or individual assessment **Domain:** Content knowledge (mathematics), pedagogical content knowledge (mathematics)
Publisher: University of Michigan School of Education Deborah Loewenberg Ball 610 E. University Ave. Ann Arbor, MI 48109 deborahball@umich.edu http://www.soe.umich.edu		**Grade/Age Range:** Kindergarten through grade 6 **Administration Interval:** None described
Material, Training, and Scoring Costs: Not specified		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2+ (certification beyond bachelor's like a master's) **Personnel for Administration:** Individual with basic clerical skills with some training **Training for Administration:** Basic test timing and proctoring
Languages: English		**Alternate Forms:** Yes, 3 parallel forms; may be administered as frequently as 3 weeks apart
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** Not available **Time to Administer:** 30 to 35 minutes **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3[1] (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Pedagogical Content Knowledge Assessment (PCK) is an individual or group administered, multiple-choice assessment that measures the content knowledge and pedagogical content knowledge of kindergarten through grade 6 teachers in mathematics. Content knowledge is defined as an understanding of mathematical material, and pedagogical content knowledge is an understanding of how students think about, know, or learn such material; it is distinct from teachers' subject matter knowledge. A team of developers began development of the PCK in 1999; the most current psychometric information comes from a 2008 report on a pilot test in 2000–2001. The knowledge of content and students is the area of PCK that is addressed, and the items fall into one of four categories: (1) common student errors, (2) students' understanding of content, (3) student developmental sequences, and (4) common student computational strategies. The paper-and-pencil assessment averages 20 items on a form (some items share a common stem and thus are not independent items) and takes an average of 30 to 35 minutes to complete.

Other Languages: None.

Uses of Information: The PCK assesses teachers' knowledge of how students think about and learn mathematics—what the authors call knowledge of content and students (KCS)—and content knowledge (CK).

Methods of Scoring: Each item is scored as correct or incorrect. Scoring is based on a multidimensional Item Response Theory (IRT) model, to obtain scores for two dimensions (KCS, CK). The developers have required scores to be reported in an IRT metric.

Interpretability: The PCK has been used with kindergarten through grade 5 teachers involved in studies of curricula or professional development in mathematics. Only an individual with expertise in psychometrics should interpret PCK results. In general, higher scores indicate a greater understanding of how children learn mathematics and how to teach mathematics to children.

Reliability:
(1) Internal consistency reliability: The authors report reliability estimates for a one-dimensional two-parameter logistic IRT model based on combined pre- and post-test data for a single sample: 0.71 for Form A; 0.73 for Form B; and 0.78 for Form C (Hill et al. 2004). In the case of a unidimensional Rasch model to estimate scores separately for each time point, reliability estimates ranged from 0.58 (Form C pre-test) to 0.69 (Form C post-test). The authors note that these reliability coefficients were lower than hoped for and hypothesized that they were indicative of a weakness in the ability of items on the assessment to discriminate between individuals at the higher end of ability. The authors state that the measurement is more precise for individuals one to two standard deviations below the average item difficulty level (the peak of the test information curve occurs within the range of -1.50 to -0.88 logits on the IRT metric).
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: No information available.
(4) Inter-rater reliability: No information available.

Validity Evidence:
PCK measures both mathematics content knowledge (CK) and knowledge of students in relation to mathematics content (KCS). In creating items for the measure, the authors drew on (1) research about how students develop mathematically, how they solve mathematics problems, and their difficulties with particular aspects of mathematics and (2) the authors' own classroom experiences. The items focus on the development and errors made by the typical student in the United States regardless of curricula, instructional methods, or other influences. The authors piloted the PCK in 2000 and 2001 as part of California's Mathematics Professional Development Institutes. They conducted cognitive interviews with teachers, focusing on six KCS items on the assessment. The results suggested remaining problems with the conceptualization of the domain and with the design of distractors for the multiple-choice format; in the latter case, teachers rarely selected the outright "wrong" answers potentially because the items "taught" content, allowing the teachers to use deductive reasoning.

Construct/Concurrent validity: Exploratory factor analysis indicated that KCS items formed their own factor, but some items loaded on the CK factor as well. Using confirmatory factor analysis, the authors determined that most items loaded on both the KCS factor and a "general" factor, suggesting that teachers used both KCS and subject matter knowledge to respond to the KCS items.

A large-scale longitudinal study of whole-school reform efforts administered a sample of 30 items drawn from the item pool (Forms A, B, and C). The items were embedded within a survey instrument administered to teachers over three years. The responses of the grade 1 and 3 teachers were analyzed with a two-parameter logistic IRT model, with the reliability coefficient for raw scores at 0.88. The teachers' content knowledge for teaching was associated with student gains on the TerraNova mathematics test, which functioned as the strongest teacher-level predictor in linear mixed models that estimated the influence of student, teacher, and school characteristics on student gain scores (Hill et al. 2005).

With respect to discriminant analysis, the authors found that teachers in professional development institutes that focused on KCS-related topics achieved higher post-test scores on the assessment. Conversely, the authors found that teachers in professional development institutes that more generally focused on subject matter knowledge failed to realize any statistically significant effect on post-test scores on the assessment.

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: No information available.

Adaptations/Special Instructions for Individuals with Disabilities: No information available.

Alternate Forms: The PCK has three parallel forms—A, B, and C—that are linked by using conventional IRT methods. The alternate forms have been used as frequently as three weeks apart.

Previous Version: None.

NCEE or REL Study Use:[2] Achievement Effects of Four Early Elementary School Math Curricula

[1] The rating reflects a two-parameter model. Using a one-parameter model demonstrated internal consistency reliability coefficients below the 0.70 level.

[2] See Table F.1 for web address.

References:

Hill, Heather C., Deborah Loewenberg Ball, and Stephen G. Schilling. "Unpacking Pedagogical Content Knowledge: Conceptualizing and Measuring Teachers' Topic-Specific Knowledge of Students." *Journal for Research in Mathematics Education*, vol. 39, no. 4, 2008, pp. 372-400.

Hill, Heather C., Brian Rowan, and Deborah Loewenberg Ball. "Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement." *American Educational Research Journal*, vol. 42, no. 2, 2005, pp. 371-406.

Hill, Heather C., Stephen G. Schilling, and Deborah Loewenberg Ball. "Developing Measures of Teachers' Mathematics Knowledge for Teaching." *Elementary School Journal*, vol. 105, 2004, pp. 11-30.

Schilling, Stephen G. "The Role of Psychometric Modeling in Test Validation: An Application of Multidimensional Item Response Theory." *Measurement: Interdisciplinary Research and Perspectives*, vol. 5, 2007, pp. 93-106.

REFORMED TEACHING OBSERVATION PROTOCOL (RTOP), 2000

Authors: Michael Piburn and Daiyo Sawada		**Type of Assessment:** Classroom observation **Domain:** Instructional practices (science and mathematics); pedagogical content knowledge (science, mathematics)
Publisher: Center for Research on Education in Science, Mathematics, Engineering and Technology (CRESMET) Evaluation Facilitation Group Arizona State University http://cresmet.asu.edu		**Grade/Age Range:** Kindergarten through graduate programs **Administration Interval:** None described
Material, Training, and Scoring Costs: Materials and training available online at no cost		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2 (bachelor's degree with coursework in measurement and domain) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours)
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (under $100) **Time to Administer:** 90 minutes **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Constructive/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Reformed Teaching Observation Protocol (RTOP) is a classroom observation measure of instructional practices and pedagogical content knowledge used to assess the degree to which mathematics and/or science instruction has been "reformed" as defined by the Arizona Collaborative for Excellence in the Preparation of Teachers (ACEPT) project[1] and other professional societies. According to ACEPT, reformed teaching is standards-based and inquiry-oriented. The RTOP focuses on student/teacher dialogue and collaboration as well as on adaptive lessons that incorporate students' diverse levels of knowledge. The RTOP is a paper-and-pencil observation that may be used in classrooms from kindergarten through graduate programs, although there is no evidence of the measure's use in elementary school classrooms. The measure comprises 25 items (rated from not observed to very descriptive) divided evenly into five categories (or subscales): (1) lesson design and implementation; (2) content related to propositional pedagogic knowledge or knowledge of "what is"; (3) content related to procedural pedagogic knowledge or knowledge of "how to"; (4) classroom culture of communicative interactions; and (5) classroom culture of student/teacher relationships.

Other Languages: None.

Uses of Information: The RTOP is used to measure the extent to which reformed teaching practices in science and mathematics are evident in classrooms. The authors also note that teachers may use the RTOP for self-assessment of their instructional practices, for mentoring and professional development purposes, and as a checklist for lesson planning.

Methods of Scoring: Observers score items on a scale of 0 (not observed) to 4 (very descriptive) as a measure of the degree to which an activity occurred, not as a measure of frequency. For some items, the authors note, a rating may be given only after observing the entire lesson; the authors encourage observers to take notes during the lesson and then provide a rating at the observation's conclusion. Total scores range from 0 to 100, with higher scores correlating with a greater degree of instructional reform. Observers also record information about the classroom, instructor, and lesson observed; the duration of the observation; the setting (space, seating arrangements); and student details (number, gender, ethnicity, and so forth).

Interpretability: The training manual provides guidance on the interpretation of each item of the RTOP. It also provides a point of reference for instructor scores based on the data the developers collected in the pilot study.

Reliability:
(1) Internal consistency reliability: With a sample of observations from more than 141 mathematics and science classrooms in middle schools, high schools, and community colleges and universities, the authors calculated a Cronbach's alpha of 0.97; the alpha for scores from the five subscales ranged from 0.80 to 0.93.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: Not applicable.
(4) Inter-rater reliability: Two observers from the Evaluation Facilitation Group (EFG) conducted 17 pairs of observations in 153 mathematics and science classrooms in middle schools, high schools, and community colleges and universities. One pair was discarded,

however, because the observers discussed the lesson, eliminating independence. Three of the pairs were observations of the same classroom but were conducted on different days; the results were included in the analyses. To estimate inter-rater reliability, the developers used a "best-fit linear regression on one set of observations versus the other." The resulting R-squared using the total RTOP score was 0.95. For the subscales using the same method, the R-squared ranged from 0.67 (propositional pedagogic knowledge) to 0.95 (procedural pedagogic knowledge). Two other members of the EFG conducted another evaluation, of eight biology instructors; the correlation coefficient was 0.90, and the R-squared resulting from the best-fit linear regression was 0.80. Previous measures of inter-rater reliability in smaller samples all exceeded 0.90.

Validity Evidence:
The RTOP draws on the principles of reform underlying the ACEPT project as well as on resources from professional organizations.[2] Items were initially drawn from two instruments: the Horizon Research 1997–1998 Local Systemic Change Revised Classroom Observation Protocol and a classroom observation measure developed by an ACEPT researcher and adapted by the developers. The EFG narrowed the item pool from 60 to the final 25 by eliminating those that did not strongly focus on reform. The developers used videotapes to adapt and refine the items in an iterative manner. The RTOP was then used in university and college classrooms in spring 1999 and piloted in fall 1999 in 153 mathematics and science classrooms in middle schools, high schools, and community colleges and universities. The authors also conducted exploratory factor analysis by using principal components analysis and found a three-factor structure explaining 71.9 percent of the variance. The first factor, inquiry orientation, comprised 19 items (from all but the content propositional pedagogical knowledge subscale); the second factor, content propositional pedagogical knowledge, comprised 5 items (exclusively from the subscale of the same name); and the third factor, collaboration, comprised 3 of the 5 items of the student/teacher relationships subscale.

Construct\Concurrent validity: Sixteen instructors in mathematics, physical science, and physics were observed in fall 1999 at least twice. The authors calculated the correlation between RTOP scores and normalized student achievement gain scores, with correlation coefficients of 0.88 for physical science (N = 6), 0.92 for number sense, 0.94 for mathematics conceptual understanding (N = 6), and 0.97 for physics (N = 4). In a study of individual college-level classrooms, the authors also calculated correlations between the instructors' mean RTOP scores and students' normalized gain achievement scores on subtests of concept understanding (0.86) and student reasoning (0.70) and correlations between instructors' mean RTOP scores and students' mean post-test scores on the number sense subtest (0.92). The authors also found significant differences between the RTOP scores of instructors with exposure to ACEPT reform training (N = 20) and instructors without ACEPT reform experience (N = 8).

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: The Evaluation Facilitation Group may be contacted to conduct training workshops. Training via the Internet is also available at no cost so that individuals may familiarize themselves with the RTOP, but Internet-based training does not provide the necessary skills to use the RTOP in research.

Adaptations/Special Instructions for Individuals with Disabilities: Not applicable.

Alternate Forms: Not applicable.

Previous Version: None.

NCEE or REL Study Use:[3] Effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI; REL-Southeast)

[1] The Arizona Collaborative for Excellence in Preparation of Teachers (ACEPT) operated from 1994-2003 as a National Science Foundation-funded project to improve undergraduate science and mathematics teaching in Arizona colleges and universities.

[2] Resources used to develop items include the National Council for the Teaching of Mathematics's *Curriculum and Evaluation Standards* (1989), *Professional Teaching Standards* (1991), and *Assessment Standards* (1995); the National Academy of Science National Research Council's *National Science Education Standards* (1995); and the American Association for the Advancement of Science Project 2061's *Science for All Americans* (1990) and *Benchmarks for Scientific Literacy* (1993).

[3] See Table F.1 for web address.

References:

Lawson, Anton, Russell Benford, Irene Bloom, Marilyn Carlson, Kathleen Falconer, David Hestenes, Eugene Judson, Michael Piburn, Daiyo Sawada, Jeff Turley, and Susan Wyckoff. "Evaluating College Science and Mathematics Instruction: A Reform Effort that Improves Teaching Skills." *Journal of College Science Teaching*, vol. 31, no. 6, 2002, pp. 388-393.

MacIsaac, Dan, and Kathleen Falconer. "Reforming Physics Education Via RTOP." *The Physics Teacher*, vol. 40, no. 8, 2002, pp. 479-485.

Piburn, Michael, Daiyo Sawada, Kathleen Falconer, Jeff Turley, Russell Benford, and Irene Bloom. "Reformed Teaching Observation Protocol--about RTOP." Available at [http://physicsed.buffalostate.edu/AZTEC/RTOP/RTOP_full/about_RTOP.html]. 2003.

Piburn, Michael, Daiyo Sawada, Kathleen Falconer, Jeff Turley, Russell Benford, and Irene Bloom. "Reformed Teaching Observation Protocol (RTOP)." No. ACEPT IN-003. Available at [http://PhysicsEd.BuffaloState.Edu/AZTEC/rtop/RTOP_full/PDF/]. 2000.

Sawada, Daiyo, Michael Piburn, Eugene Judson, Jeff Turley, Kathleen Falconer, Russell Benford, and Irene Bloom. "Measuring Reform Practices in Science and Mathematics Classrooms: The Reformed Teaching Observation Protocol." *School Science and Mathematics*, vol. 102, no. 6, 2002, pp. 245-253.

TEST OF ECONOMIC LITERACY, THIRD EDITION (TEL-3), 2001

The Test of Economic Literacy, Third Edition (TEL-3) has a student version and thus is included in Appendix B, Student Achievement/Development Measures. Please refer to Appendix B for this profile.

TABLE C.1

NCEE OR REL RECENTLY DEVELOPED TEACHER KNOWLEDGE MEASURES SUMMARY

Measure	NCEE or REL Study (source)[a]	Description	Grade/Age Range[b]	Reliability	Validity
Reading Knowledge (content and/or pedagogical)					
Reading Content and Practices Survey (RCPS)	Professional Development Interventions on Early Reading (1)	This measure is an individual assessment of subject knowledge and pedagogical content knowledge (PCK) pertaining to reading. In particular, the measure features 30 items covering content knowledge and PCK in five areas: (1) phonemic awareness, (2) phonics, (3) fluency, (4) vocabulary, and (5) comprehension skills. The items in each area differ in the balance of content knowledge and PCK, with the first two areas having most items on content, the second two an equal balance, and the fifth area comprising mostly PCK items. Three scores are generated across both content and PCK domains: (1) a total score on overall knowledge of reading and its practices; (2) a word-level subscale score measuring knowledge in the first three areas (phonemic awareness, phonics, and fluency, representing 50 percent of the items in each form); and (3) a meaning-level subscale score measuring the latter two areas (vocabulary development and reading comprehension, representing 50 percent of the items in each form). Teachers complete one of six versions of the assessment, each with 30 items drawn from a bank of 90 items. For each version, 27 to 29 of the 30 items are multiple-choice questions, with the remainder short-answer questions. The	Grade 2	Rasch (IRT) reliability, presented separately for the implementation (I) and follow-up (F) study years: Total scale: 0.60 (I), 0.56 (F) Word-level subscale: 0.45 (I), 0.46 (F) Meaning-level subscale: 0.49 (I), 0.42 (F)	A confirmatory factor analysis demonstrating a two-factor model (word level and meaning level) fit the data significantly better than a single-factor model. Pearson's *r* correlations between the word and meaning subscale scores were 0.38 and 0.32, respectively, across the implementation and follow-up years of the study. Correlations corrected for measurement error are also presented.

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.

[b]Study sample was used to determine grade/age range of measure.

Table C.1 *(continued)*

Measure	NCEE or REL Study (source)[a]	Description	Grade/Age Range[b]	Reliability	Validity
		measure requires 30 minutes for completion. Scores are based on a Rasch model. Approximately 270 teachers from 90 schools in six districts comprised the sample.			
Teacher impact questionnaire of ELL instructional pedagogy	Principles-Based Professional Development (REL-Pacific) (3)	This measure is an individual assessment of content knowledge and pedagogical knowledge, measuring teachers' understanding of appropriate pedagogy for teaching English Language Learners (ELL). The questionnaire contains approximately 46 multiple-choice questions in two categories: instructional techniques (e.g., identifying the definition of schema building) and theories of learning and language acquisition (e.g., identifying the definition of additive bilingualism). The questionnaire requires about 45 minutes for completion. Approximately 270 teachers from 50 schools are expected to complete the questionnaire.	Grades 4–5	Not available	Not available

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.

Table C.1 *(continued)*

Measure	NCEE or REL Study (source)[a]	Description	Grade/Age Range[b]	Reliability	Validity
Mathematics Knowledge (content and/or pedagogical)					
Teacher Knowledge Inventory (TKI)	Professional Development Strategies in Math (3)	This measure is an individual assessment of subject knowledge and pedagogical content knowledge (PCK) pertaining to mathematics. The TKI consists of three forms, each with 24 items. The items are divided equally between common content knowledge (CCK) and PCK. CCK items measure conceptual knowledge related to a specific key understanding (such as fractions) while PCK items measure content-specific pedagogical skills in the areas of planning, delivering, and assessing instruction in relation to the key concepts. The CCK items are multiple-choice and short-constructed-response questions; the PCK items are all multiple-choice. The 24 items are also divided equally across 12 key content areas: 6 items for fractions and decimals and 6 items for ratios, proportions, and percents. The TKI is designed to be administered via a paper questionnaire in 45 minutes. The study plans for approximately 214 teachers from 84 schools to complete the TKI.	Grade 7	Not available	External mathematicians reviewed the TKI. Each item was also pretested in cognitive think-aloud interviews with at least six teachers, and the entire TKI was piloted in proctored small-group sessions.

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.

Table C.1 *(continued)*

Measure	NCEE or REL Study (source)[a]	Description	Grade/Age Range[b]	Reliability	Validity
Pedagogical Knowledge					
Test of assessment knowledge	Classroom Assessment for Student Learning (REL-Central) (3)	This measure is an individual assessment of teacher pedagogical knowledge on classroom assessment. It is an online assessment that contains multiple-choice and true-false questions on the knowledge and reasoning skills taught by the Classroom Assessment for Student Learning program (e.g., asking the teacher to identify the most appropriate form of assessments for different instructional goals) as well as on general classroom practices (e.g., identifying strategies that can improve class discussions) and assessment (e.g., identifying a use of formative assessment). The study team developed a pool of 72 items for the assessment, with approximately 50 items planned for the final baseline measure. Follow-up administrations may contain fewer items based on the reliability and validity of the items at baseline. Approximately 265 teachers in 64 schools will complete the assessment.	Grades 4–5	Not available	Not available

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.

RECENTLY DEVELOPED TEACHER MEASURES REPORT REFERENCES

Garet, Michael S., Stephanie Cronen, Marian Eaton, Anja Kurki, Meredith Ludwig, Wehmah Jones, Kazuaki Uekawa, Audrey Falk, Howard Bloom, Fred Doolittle, Pei Zhu, and Laura Sztejnberg. "The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement." (NCEE 2008–4030). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, September 2008.

APPENDIX D

CLASSROOM PRACTICES AND SETTINGS MEASURE PROFILES AND TABLE OF RECENTLY DEVELOPED MEASURES

AUTHENTIC INSTRUCTIONAL PRACTICES
CLASSROOM OBSERVATION FORM, 1993

Authors: Jerome D'Agostino; based on observational framework of Fred M. Newmann and Gary G. Wehlage		**Type of Assessment:** Classroom observation **Domain:** Instructional practices (comprehensive, reading, math), classroom quality (teacher-student interactions), school engagement
Publisher: Not commercially published; see D'Agostino (1996)		**Grade/Age Range:** Kindergarten through Grade 12 **Administration Interval:** As frequently as desired
Material, Training, and Scoring Costs: Not specified		**Personnel and Training Requirements** **Credentials Required for Use:** Not specified **Personnel for Administration:** Not specified **Training for Administration:** Not specified While personnel training is not specified, observers need to make complex distinctions on pedagogy that require background knowledge.
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (<$100) **Time to Administer:** Duration of observed lesson **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Authentic Instructional Practices Classroom Observation Form (D'Agostino 1996) is an adapted version of Newmann and Wehlage's (1993) observational framework for assessing authentic instruction standards (i.e., instruction promoting higher-order thinking skills, depth of knowledge, connectedness to students' lives outside the classroom, substantive conversation, and social support to advance student achievement) in elementary through secondary classrooms. D'Agostino (1996) adapted the framework to assess authentic instruction in mathematics and reading for grade 3 students and investigated the validity of results. The form consists of items assessing higher-order thinking skills, coherence of subject matter, connection to students' out-of-school experiences, substantive conversation, social support, and student engagement. For the higher-order thinking items (six mathematics items and seven reading items), observers rate on a four-point scale how central an activity or concept is to the substance of a lesson (0 = none, 1 = low emphasis, 2 = moderate emphasis, 3 = high emphasis). The other items have item-specific descriptors associated with increasing amounts of an activity or concept along a four-point scale (for example, for Coherence of Material, 0 = "Material is presented in superficial fragments with very little connection between parts"; 3 = "Key concepts/ideas are covered in depth. The lesson content is presented as a whole, and is structured in a way that allows for the sequencing and structuring of a complex topic. Each topic appears to build on another in an effort to foster deeper student understanding."). The same descriptors apply to both mathematics and reading items, and observers select the number that most accurately describes the mathematics or reading lesson. According to D'Agostino (1996), all of the items together comprise "Authentic Instruction factors for both math and reading" (p. 143).

More recently, Borman et al. (2000) used a modified version of D'Agostino's (1996) observation form to assess authentic instruction in kindergarten through grade 8 Title I classrooms. In contrast to D'Agostino's (1996) version, Borman et al.'s (2000) adaptation rates instruction across all academic subject areas (not just mathematics and reading). The higher-order thinking skills items used by Borman et al. (2000) were not mathematics- and reading-specific but are based on Bloom's (1956) taxonomy: knowledge, comprehension, application, analysis, synthesis, and evaluation.

Other Languages: None.

Uses of Information: As a research tool, the observational framework may be used to estimate levels of authentic instruction in elementary, middle, and high schools. It assesses how levels and qualities of authentic instruction relate to student achievement, organizational features of schools, educational programming, school leadership, and school and community culture. It may also be used to study how school reforms influence instruction. In addition, the framework may find application as a professional development tool for individual teachers or groups of teachers.

Methods of Scoring: Specific scoring instructions are not available. Borman et al. (2000) standardized the items and calculated mean scores. D'Agostino (1996) combined ratings on the higher-order thinking skills (HOTS) items to produce an overall score for those items for each classroom. Based on classroom mean scores on those items, he categorized rooms into one of four groups (derived from breaks in the frequency distributions of the means). He assigned a rating from zero for classrooms with the lowest mean scores to three for classrooms at the top of

the distribution for HOTS. He performed the procedure separately for mathematics and reading lessons to generate overall higher-order thinking skills scores for both lesson types. D'Agostino used multifaceted Rasch rating scale models to estimate scores for two constructs, Authentic Reading Instruction (ARI) and Authentic Math Instruction (AMI), using the HOTS categories and the other five ratings associated with the respective content area.

Interpretability: Higher mean ratings indicate greater emphasis on higher-order thinking and other authentic instructional practices. No guidelines are presented on interpreting scores.

Reliability:
(1) Internal consistency reliability: Borman et al. (2000) reported an alpha coefficient of 0.82 for scores from their version of the measure. D'Agostino (1996) reported Item Response Theory reliability of 0.84 for both the ARI and the AMI.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: Not applicable.
(4) Inter-rater reliability: Newmann and Wehlage (1993) reported inter-rater reliability coefficients of 0.70 or higher, with precise agreement between observers at or above 60 percent or higher for each standard. D'Agostino (1996) estimated each of five observers' severity levels based on paired observations conducted in 10 classrooms. Multifacted Rasch analysis yielded fit statistics for each observer that indicated inter-rater agreement or non-agreement. For observer ratings of both the mathematics and the reading lessons, the author reported good rater fit, though observers differed significantly in the severity of their ratings, particularly in the case of a single observer on the AMI. The multifaceted Rasch model adjusted classroom scores for the differences in severity among observers. D'Agostino used the fit statistic as the assessment of inter-rater reliability. He reported that fit was adequate because none of the mean square outfit values for observers was greater than 1.3 for either mathematics or reading.

Validity Evidence:
Newmann and Wehlage's (1993) framework assesses five standards for authentic instruction that the authors theorized would improve instruction and student achievement in elementary through secondary schools (see Description). Later adaptations reflected the same theoretical foundation.

D'Agostino (1996) collected validity evidence on his version of the measure with a sample of 53 randomly chosen, self-contained Title I grade 3 classrooms in 29 schools in the Chicago Public Schools. The research team observed one mathematics and one reading lesson in 52 classrooms. In the 53rd classroom, only a mathematics lesson was observed. Mathematics and reading lessons lasted an average of 49 and 61 minutes, respectively. In all of the schools, more than 60 percent of the students lived in poverty. In 25 of the schools, at least 90 percent of the students were Black; 4 schools served mostly Hispanic children. The average class size was 15.5 students, and 79 percent of classes were held in regular classrooms. The validity study was conducted during the 1993–1994 school year.

Construct/Concurrent validity: Using multifaceted Rasch models, D'Agostino (1996) investigated the construct validity of two latent traits: the Authentic Reading Instruction (ARI) and Authentic Math Instruction (AMI). Items on the ARI and the AMI showed adequate item fit, and the item separation reliability estimates for both were 0.95. The majority of classrooms

exhibited good fit with the model, with classroom separation reliability estimates of 0.84 for both the ARI and the AMI.

D'Agostino (1996) hypothesized that higher levels of authentic instruction would be linked to gains in mathematics skills (computation and problem solving) and student reading skills (vocabulary and comprehension) as assessed with the subtests of the Iowa Test of Basic Skills. He examined a series of two-level hierarchical linear models (HLM) (students within classrooms) with adjusted gain scores (the difference between a post-test score and a predicted post-test score based on a regression of the post-test scores on the pre-test scores) as the outcome and classroom observation scores as the only predictor. Intraclass correlations showed that 28 to 40 percent of the variance in students' adjusted gains was at the classroom level. The model for vocabulary indicated that the ARI did not predict mean classroom gains. A similar model for reading comprehension supported a non-linear quadratic association, with the ARI predicting increased mean classroom gains up to the mean ARI and then leveling off. The quadratic model accounted for 8 percent of the classroom-level variance in reading comprehension gains. With respect to mathematics instruction, HLM models indicated that the AMI significantly predicted classroom mean gains in both computation and problem-solving, accounting for approximately 13 and 9 percent of classroom-level variance, respectively.

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: No information available.

Adaptations/Special Instructions for Individuals with Disabilities: Not applicable.

Alternate Forms: Not applicable.

Previous Version: D'Agostino (1996) adapted the Authentic Instructional Practices Classroom Observation Form from Newmann and Wehlage's (1993) observational framework for assessing authentic instruction with secondary students, which had not been validated empirically. His adaptation assesses authentic instruction in mathematics and reading by using item descriptors appropriate for grade 3 students. It also uses a four-point scale response format in contrast to the five-point scale used by Newmann and Wehlage (1993). Borman et al. (2000) used an adaptation of D'Agostino's (1996) form with students in kindergarten through grade 8.

NCEE or REL Study Use:[1] The Effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI) (REL-Southeast)

[1] See Table F.1 for web address.

References:

D'Agostino, Jerome V. "Authentic Instruction and Academic Achievement in Compensatory Education Classrooms." *Studies in Educational Evaluation*, vol. 22, no. 2, 1996, pp. 139-155.

Bloom, Benjamin S. (ed.). *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*. New York: Longmans, 1956.

Borman, Geoffrey D., Laura Rachuba, Amanda Datnow, Marty Alberg, Martha MacIver, Sam Stringfield, and Steve Ross. "Four Models of School Improvement: Successes and Challenges in Reforming Low-Performing, High-Poverty Title I Schools." No. 48. Baltimore: Center for Research on the Education of Students Placed At Risk, 2000.

Newmann, Fred M., and Gary G. Wehlage. "Five Standards of Authentic Instruction." *Educational Leadership*, vol. 50, no. 7, 1993, pp. 8-12.

CAREGIVER INTERACTION SCALE (CIS), 1989

Authors: Jeffrey Arnett		**Type of Assessment:** Classroom observation **Domain:** Classroom quality (teacher-student interactions)
Publisher: Not commercially available. A copy of the scale may be found in Jaeger and Funk (2001).		**Grade/Age Range:** Caregivers/teachers of preschool-age children **Administration Interval:** None described
Material, Training, and Scoring Costs: None		**Personnel and Training Requirements** **Credentials Required for Use:** Level 1 (training or supervised experience with measurement) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours) To be considered a certified Caregiver Interaction Scale observer, one's observation scores should demonstrate a 0.70 inter-rater reliability coefficient for two consecutive visits (Jaeger and Funk 2001).
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: None described		**Summary** **Initial Material Cost:** 1 (<$100) **Time to Administer:** 90 minutes or more **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The 26-item Caregiver Interaction Scale assesses the quality and content of a teacher's interactions with preschool-age students. The scale was designed to provide information on socialization practices identified in research on parenting, with items based on observations of Head Start teachers. The scale may be used without modification in both center and home-based settings. The items measure the emotional tone, discipline style, and responsiveness of the caregiver or teacher in the classroom. The items are organized into the following four subscales: (1) positive interaction (10 items: warm, enthusiastic, and supports developmentally appropriate behavior); (2) punitiveness (9 items: hostility, harshness, and use of threat); (3) detachment (4 items: uninvolvement and disinterest); and (4) permissiveness (3 items: fails to supervise, low on control). Observers rate the items for the extent exhibited from "not at all" to "very much". The observation time varies, but Arnett (1989) observed caregivers in two 45-minute sessions while Jaeger and Funk observed caregivers in a 2.5-hour session.

Other Languages: None.

Uses of Information: The scale may be used to assess caregivers' interactions with students and their emotional tone and approach to engaging and disciplining students.

Methods of Scoring: Observers rate the extent to which the caregiver or teacher exhibits the behavior described in each item. The rating falls on a four-point scale: "not at all" (1), "somewhat" (2), "quite a bit" (3), or "very much" (4). The total score may be calculated by summing the ratings across the 26 items. Averages may be calculated for each subscale.

Interpretability: Depending on the school or classroom's needs, individual caregiver scores may be compared to the scores of other caregivers, or the mean scores of a group of caregivers may be compared to the means of other groups of caregivers. Those interpreting the scale often use statistical tests to assess the differences between scores.

Reliability:
(1) Internal consistency reliability: Layzer et al. (1993) obtained Cronbach's alphas of 0.91 for scores of the warmth/responsiveness (positive interaction) subscale and of 0.90 for scores of the harshness (punitiveness) subscale while Resnick and Zill (1999) obtained Cronbach's alphas of 0.98 and 0.93 for total scores across all items of, respectively, Head Start lead teachers and assistant teachers. Jaeger and Funk (2001) reported reliability coefficients of 0.81 and higher for scores of the sensitivity (positive interaction), punitiveness, and detachment subscales.
(2) Test-retest reliability: No information available.
(3)Alternate form reliability: Not applicable.
(4) Inter-rater reliability: Jaeger and Funk (2001) reported inter-rater reliability coefficients ranging from 0.75 to 0.97 between a certified observer and trainees across two observations.

Validity Evidence:
Arnett developed items on the CIS during a pilot study conducted in Head Start centers in Charlottesville, Virginia, with caregivers of preschool children. Items were tested until 80 percent agreement was achieved among three observers involved in the pilot study (Arnett 1989). Subsequent factor analyses by Arnett revealed four primary factors that match the subscales:

positive interaction, punitiveness, detachment, and permissiveness. Additional factor analyses conducted in further studies show either three or four factors: (1) sensitivity, (2) harshness (punitiveness), and (3) detachment (Howes et al. 1989) versus (1) attentive and encouraging, (2) harsh and critical, (3) detached, and (4) controlling (Love et al. 1992).

Construct/Concurrent validity: Layzer et al. (1993) reported correlation coefficients from 0.43 to 0.67 between the CIS and the Early Childhood Environment Rating Scale (ECERS), the Assessment Profile for Early Childhood Programs, and the Description of Preschool Practices. The authors stated they did not expect the coefficients to be large because the CIS scale focuses narrowly on an aspect of teacher behavior that is not directly measured by the other three observation instruments. In another study, a correlation of 0.76 was found between the CIS and the ECERS (Phillipsen et al. 1995).

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: No information available.

Adaptations/Special Instructions for Individuals with Disabilities: Not applicable.

Alternate Forms: Not applicable.

Previous Version: None.

NCEE or REL Study Use:[1] Evaluating the Impact of the Program for Infant/Toddler Care

[1] See Table F.1 for web address.

References:

Arnett, Jeffrey. "Caregivers in Day-Care Centers: Does Training Matter?" *Journal of Applied Developmental Psychology*, vol. 10, 1989, pp. 541-552.

Halle, Tamara, and Jessica Vick. "Quality in Early Childhood Care and Education Settings: A Compendium of Measures." Submitted to the Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Washington, DC: Child Trends, November 2007.

Jaeger, Elizabeth, and Suzanne Funk. "The Philadelphia Child Care Quality Study: An Examination of Quality in Selected Early Education and Care Settings." Philadelphia: Saint Joseph's University, 2001.

Kisker, Ellen E., Kimberly Boller, Charles Nagatoshi, Christine Sciarrino, Vinita Jethwani, Teresa Zavitsky, Melissa Ford, and John M. Love. "Resources for Measuring Services and Outcomes in Head Start Programs Serving Infants and Toddlers." Washington, DC: Mathematica Policy Research, 2003.

Layzer, Jean I., Barbara D. Goodson, and Marc Moss. "Observational Study of Early Childhood Programs, Final Report, Volume I: Life in Preschool." Cambridge, MA: Abt Associates, Inc., 1993.

Love, John M., Alicia Meckstroth, and Susan Sprachman. "Measuring the Quality of Program Environments in Head Start and Other Early Childhood Programs: A Review and Recommendations for Future Research." Working Paper No. 97-36. Washington, DC: National Center for Education Statistics, U.S. Department of Education, 1997.

Love, John M., Paul Ryer, and Bonnie Faddis. "Caring Environments: Program Quality in California's Publicly Funded Child Development Programs: Report on the Legislatively Mandated 1990–91 Staff/Child Ratio Study." Portsmouth, NH: RMC Research Corporation, 1992.

Phillipsen, Leslie, Debby Cryer, and Carollee Howes. "Classroom Process and Classroom Structure." In *Cost, Quality, and Child Outcomes in Child Care Centers*, edited by Suzanne Helburn. Denver: Department of Economics, Center for Research in Economics and Social Policy, University of Colorado at Denver, 1995.

Resnick, Gary, and Nicholas Zill. "Is Head Start Providing High-Quality Education Services? 'Unpacking' Classroom Processes." Paper presented at the biennial meeting of the Society for Research in Child Development. Albuquerque, April 1999.

Whitebook, Marcy, Carollee Howes, and Deborah Phillips. "Who Cares? Child Care Teachers and the Quality of Care in America. Final Report of the National Childcare Staffing Study." Oakland, CA: Childcare Employee Project, 1989.

CIERA CLASSROOM OBSERVATION SCHEME FOR CLASSROOM LITERACY INSTRUCTION, 2000

Authors: Barbara Taylor and P. David Pearson		**Type of Assessment:** Classroom observation **Domain:** Classroom quality (teacher-student interactions, classroom environment), instructional practices (reading), school engagement
Publisher: Center for the Improvement of Early Reading Achievement (CIERA) University of Michigan School of Education http://www.ciera.org		**Grade/Age Range:** Kindergarten through grade 6 **Administration Interval:** As frequently as desired
Material, Training, and Scoring Costs: Contact authors for information about availability and costs of observation forms, codebook, and training materials.[1] Coding categories and codes are listed in Taylor et al. (2005).		**Personnel and Training Requirements** **Credentials Required for Use:** Not specified **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (> 2 hours)
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** To be determined upon negotiation with the publisher **Time to Administer:** Duration of a reading lesson **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3[2] (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Center for the Improvement of Early Reading Achievement (CIERA) classroom observation scheme assesses seven aspects (referred to as levels) of elementary classrooms' literacy instruction: Level 1—who is delivering instruction; Level 2—grouping practices (e.g., whole class, small group); Level 3—major literacy activities or events (e.g., reading text, phonics work, discussing a story); Level 4—specific skill focus (e.g., reading text, listening to text, vocabulary, basic comprehension skills, higher-level comprehension strategies, writing, phonics); Level 5—materials (e.g., textbook, trade book, worksheet); Level 6—teacher interaction styles (e.g., telling, recitation, coaching); and Level 7—expected student responses to literacy events (e.g., reading, talking, listening). The scheme also assesses student involvement during literacy lessons by reporting the number of students observed to be on task out of all children in a class. Observers perform the observations during a series of 5-minute intervals during a reading lesson (the developers used the measure to conduct one-hour observations of reading lessons). The observer begins the first interval at or near the beginning of the lesson; at the end of the 5 minutes, the observer takes a minute or so to record codes on the coding sheet and then begins a new 5-minute interval. During each interval, the observer takes running notes on what the teacher and students say and do during the lesson. At the end of the 5 minutes, the observer records the percentage of students observed to be "on task" and codes for the literacy events (Level 3) that occurred during the interval. For each literacy event, the observer records codes for the other levels. In most cases, codes correspond to the first letter of the word(s) of the coding category (e.g., "c" for classroom teacher, "sw" for sight words; see Taylor et al. 2005). A codebook, previously available from CIERA,[1] includes instructions for assigning codes. In CIERA studies, observers also completed a post-observation summary that included open-ended questions about overall impressions and emphases, instructional practices and methods, grouping practices, details about curricular materials (e.g., book titles), student participation and engagement, classroom management, and classroom environment. The form required 15 to 20 minutes to complete.

Other Languages: None.

Uses of Information: Researchers and evaluators may use the results of observations conducted with the CIERA classroom observation scheme to assess the types and quality of literacy instruction in elementary classrooms. The results may also provide the basis for comparing instruction in different classrooms, investigating the effectiveness of instructional practices, and, if observations are collected longitudinally, assessing progress. The measure may help teachers build on strengths and identify instructional areas for further development.

Methods of Scoring: For each level, observers tabulate and record totals for each coding category within that level. In previous studies, researchers used the totals to calculate the percentage of intervals out of all observed intervals or out of a subset of intervals for a given code (Taylor et al. 2005). For example, the authors calculated percentages of intervals out of all intervals for which they observed the Level 2 codes of whole group and small group. In another example, they calculated percentages of intervals for each Level 4 code out of the number of intervals for which reading was the instructional activity. Similarly, they computed percentages of active student responses and passive student responses out of the total number of Level 7

student responses (see Taylor et al. 2005). They also calculated the mean percentage of students observed to be on task for all observed intervals.

Interpretability: The CIERA observation scheme yields quantitative and qualitative information about the nature and quality of classroom reading instruction. Quantitative data (described in Methods of Scoring) indicate frequencies of observed grouping practices, activities, interactions, materials, and student responses during instruction. Qualitative data, derived from information recorded on the post-observation summary form, consist of anecdotal accounts of instructional materials, practices, and interactions and of summary descriptions of what occurred during the observation. The developers do not provide specific guidelines for interpreting results. However, referring to the research literature on effective reading instruction, they suggest that greater frequency of empirically-proven practices is desirable. Research has supported the value of direct instruction in phonemic awareness, phonics, vocabulary, and comprehension strategies (National Reading Panel 2000). Studies have also emphasized the benefits of high levels of active student participation and engagement and of teachers' use of small-group instruction, coaching, modeling, explanation, and higher-level questioning to promote higher-level thinking (reviewed in Taylor et al. 2005).

Reliability:
(1) Internal consistency reliability: No information available.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: Not applicable.
(4) Inter-rater reliability: Taylor et al. (2002; 2005) reported percentages of agreement between two observers for each of Levels 2 through 7. The percentages of agreement ranged from 82 to 95 percent (7 of the 12 reported percentages were greater than or equal to 85 percent). In addition, both studies involved experts' subsequent rating of random samples of 10 percent of all observations. Percentages of agreement for each level between an observer and expert ranged from 82 to 98 percent (9 of the 12 percentages of reported agreement were greater than or equal to 85 percent). Percentages of agreement for each level between two expert observers ranged from 91 to 98 percent and 86 to 99 percent in the two studies, respectively (Taylor et al. 2002; 2005).

Validity Evidence:
The observation scheme was developed for use in CIERA's studies of school reform strategies for improving reading achievement among readers in high-poverty elementary schools. The coding scheme assesses key elements that had been empirically linked with effective reading instruction (reviewed in Taylor et al. 2005), such as maintaining an academic focus, clarifying learning goals, monitoring understanding, modeling, explanation, coaching (versus telling), providing feedback, emphasizing higher-level thinking skills, balancing skills-based and holistic instruction, and emphasizing small-group instruction. They also included the degree to which classroom settings were warm, democratic, cooperative, and conducive to student engagement. CIERA used the scheme in a series of studies conducted from 1997 to 2001 in urban, small town, and rural schools serving low-income students in kindergarten through grade 6. Sample sizes ranged from 92 to 134 teachers/classrooms per study.

Construct/Concurrent validity: CIERA research studies have investigated relationships between some reading instruction practices (as measured by the CIERA observation scheme) and student

reading comprehension and fluency scores on the Gates-MacGinities Reading Test and writing skills as assessed with a rubric. Taylor et al. (2005) performed three-level hierarchical linear modeling (HLM) analyses (nesting students within classrooms and classrooms within schools) with a sample of 733 grade 2 through 5 students of 92 teachers in 13 schools serving low-income students. The authors estimated three models, one for each of the dependent variables (standardized reading comprehension, reading fluency, writing skills scores). In each model, classroom-level predictor variables included selected Level 2 through Level 7 categories from the observation scheme while school-level predictor variables included school effectiveness and reform-effort scores. The analyses indicated that between-classroom variance accounted for 19 to 32 percent of the variance in student reading comprehension, fluency, and writing outcomes. Student grade level and the Level 4 category of comprehension skill instruction (rote skill work involving lower-level thinking) both related negatively to student reading comprehension scores; together, they accounted for 29 percent of between-classroom variance. Two Level 4 categories—comprehension skill instruction and high-level questioning—related negatively and positively, respectively, to student fluency and together accounted for 15 percent of between-classroom variance. With respect to student writing scores, the Level 6 category of coaching (positively related) accounted for 11 percent of between-classroom variance. These findings—that high-level questioning and coaching related positively to student reading and writing scores and that rote comprehension skill practice related negatively to student reading and writing scores—was noted as consistent with earlier research findings (Taylor et al. 2000; 2002; 2003).

An earlier study employing two-level HLM analyses (nesting students within classrooms) found significant relationships between other observation scheme categories with student reading and writing outcomes (Taylor et al. 2002). Separate analyses conducted for students in different grade ranges showed that, across analyses, between-classroom factors accounted for 21 to 49 percent of the variance in spring student reading and writing scores (after accounting for fall scores). Relationships between the Level 2 categories of small-group and whole-group instruction and student outcomes varied by grade. In grade 1, greater amounts of small-group instruction were associated with increased student fluency scores; in grades 4 through 6, whole-group instruction related positively to reading comprehension scores. With respect to Level 4 categories, highly teacher-directed practices and phonics instruction related negatively to student fluency scores (grades 2 and 3), and coaching students in word recognition strategies related positively to writing scores (grades 4 through 6). In addition, active student responses (Level 7) related positively to fluency scores (grades 4 through 6).

Finally, Taylor et al.'s (2005) HLM analyses indicated that teachers in schools that were highly engaged in school reform efforts were observed to increase practices emphasized in the reform model, in particular the use of coaching, higher-level questioning, and modeling assessed by the CIERA observation, whereas teachers in "low-reform" schools did not.

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: A classroom observation training kit consisted of a manual, kindergarten and grade 3 practice tapes, inter-rater tape for establishing inter-rater reliability, and an instructional

CD. For information about the current availability of the codebook and training kit, see http://www.ciera.org or contact the authors (contact information available on web site).

Adaptations/Special Instructions for Individuals with Disabilities: Not applicable.

Alternate Forms: Not applicable.

Previous Version: None.

NCEE or REL Study Use:[3] Impact of the Thinking Reader Software Program on Grade 6 Reading Comprehension, Vocabulary, Strategies, and Motivation; Efficacy of Frequent Formative Assessment for Improving Instructional Practice and Student Performance, Given Variations in Training to Use Assessment Results

[1] For information about the current availability of the codebook and training kit, see http://www.ciera.org or contact the authors (contact information available on web site).

[2] Percentages of agreement between two expert observers' ratings on each level met or exceeded the minimum acceptable level of 85 percent. In cases where one or both observers were not experts, some percentages of agreement were below 85 percent.

[3] See Table F.1 for web address.

References:

National Reading Panel. *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction*. Washington, DC: National Institute of Child Health and Human Development, 2000.

Taylor, Barbara. "Improving Classroom Reading Instruction: Reflecting on our Practice. CIERA Summer Institute." Ann Arbor, MI: Center for the Improvement of Early Reading Achievement. Available at [http://www.ciera.org/library/presos/2000/2000-CSI/btaylor/csi-taylor-ici.pdf], August 2000.

Taylor, Barbara, P. David Pearson, Debra Peterson, and Michael Rodriguez. "The CIERA School Change Framework: An Evidence-Based Approach to Professional Development and School Reading Improvement." *Reading Research Quarterly*, vol. 40, no. 1, 2005, pp. 40-69.

Taylor, Barbara, P. David Pearson, Debra Peterson, and Michael Rodriguez. "Reading Growth in High-Poverty Classrooms: The Influence of Teacher Practices that Encourage Cognitive Engagement in Literacy Learning." *Elementary School Journal*, vol. 104, no. 1, 2003, pp. 3-28.

Taylor, Barbara, Debra Peterson, Michael Rodriguez, and P. David Pearson. "The CIERA School Change Project: Supporting Schools as They Implement Home-Grown Reading Reform." No. 2-016. Ann Arbor, MI: Center for the Improvement of Early Reading Achievement, 2002.

The Northeast and Islands Regional Educational Laboratory. "OMB Attachment 3: CIERA Classroom Observation Protocol (Adapted for Thinking Reader)." Draft. No. (03330)1850-0837-v.1. Available at [http://edicsweb.ed.gov/browse/downldatt.cfm?pkg_serial_num=3330]. June 2007.

DIAGNOSTIC CLASSROOM OBSERVATION TOOL (DCO), 2008

The Diagnostic Classroom Observation Tool (DCO) is also a teacher knowledge measure and thus is found under Appendix C, Teacher Knowledge Measures. Please refer to Appendix C for this profile.

EARLY CHILDHOOD ENVIRONMENT RATING SCALE—REVISED EDITION (ECERS-R), 1998

Authors: Thelma Harms, Richard M. Clifford, and Debby Cryer		**Type of Assessment:** Classroom observation with some teacher report **Domain:** Classroom quality (classroom environment with some teacher-student interaction)
Publisher: Teachers College Press 800-575-6566 http://www.teacherscollegepress.com		**Grade/Age Range:** Age 2.5 through 5 years **Administration Interval:** As frequently as desired
Material, Training, and Scoring Costs: ECERS-R rating scale in spiral binder with Expanded Score Sheet and Profile for photocopying: $19.95 Video Observation for the ECERS-R and Instructor's Guide: $59 Video Guide and Training Workbook: $4 ECERS-R training (excluding travel) at the University of North Carolina (UNC): Ranges from $825 to $1,300 depending on the focus of the training		**Personnel and Training Requirements** **Credentials Required for Use:** Level 1 (training or supervised experience with measure) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours) The authors recommend that observers attend a training session (with one or more practice observations) led by an experienced ECERS-R trainer before using the scale in the field. Researchers should contact the authors regarding training for inter-rater reliability with the authors. Observers should also have knowledge of child development and educational implications (Frank Porter Graham Child Development Institute 2005).
Languages: English; French, German, Norwegian, and Spanish; Hungarian for research purposes only		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (under $100) **Time to Administer:** Administration time ranges from 2 to 5 hours depending on the scoring option **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Available[1] **Construct/Concurrent Validity:** Not available[2] **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The ECERS-R is a classroom assessment tool designed to measure the quality of group programs for students (2.5 through 5 years) in preschool, kindergarten, and child care classrooms. It is a 43-item rating scale organized into seven environmental subscales: (1) Space and Furnishings, (2) Personal Care Routines, (3) Language-Reasoning, (4) Activities, (5) Interaction, (6) Program Structure, and (7) Parents and Staff. Each item has a number of quality indicators, with 470 yes/no indicators in total. Observations take place during "play/learning times and routines, such as meal, toileting, and preparation for nap" (Halle and Vick 2007). The observer must also set aside time to speak with staff regarding unobserved indicators. Administration time varies with the scoring option and whether an outside observer is used, although administration time averages 3.5 hours.

Other Languages: The ECERS-R has been translated into other languages with the basic scale remaining the same. It is available for purchase in French, German, Norwegian, and Spanish as well as in Hungarian for research purposes only. Changes were made to certain indicators and to the examples used to describe indicators in order to reflect the culture. While the indicators underwent modification, the authors indicated the translations possess the same scale structure but did not provide information on investigations of comparability on subscale scores between the English version and other translations.

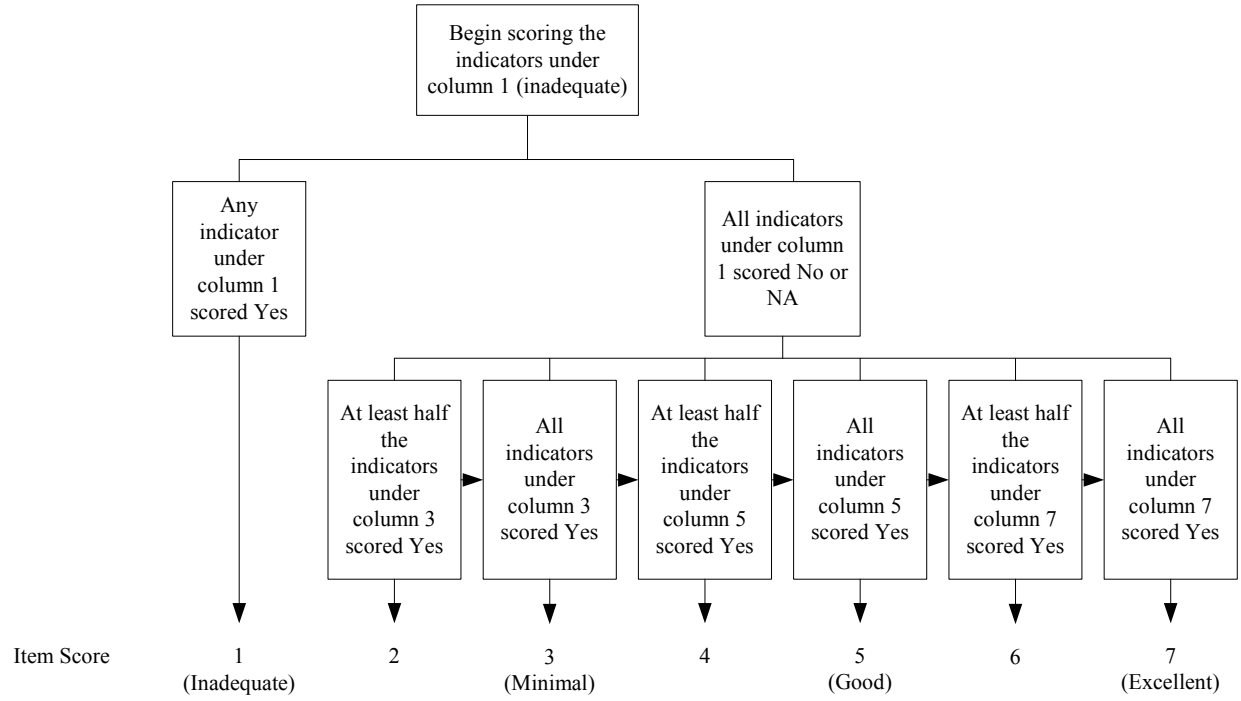
Uses of Information: The assessment may be used for research as well as by program directors for supervision and program improvement, by teaching staff for self-assessment, by agency staff for program monitoring, and by teacher training programs for instruction.

Methods of Scoring: An individual thoroughly familiar with the ECERS-R should score the assessment. The Expanded Score Sheet is used to record the ratings for quality indicators, items, subscale scores, and total scores as well as any observer comments. The indicators, which have Yes/No/Not Applicable (NA) response choices, are used to score the items from 1 (Inadequate) to 7 (Excellent). Indicators fall under columns at the scale anchors 1, 3, 5, and 7. Items may be scored two ways as described in detail in the manual. Under the standard scoring option for each item, if any of the indicators in the Inadequate column (or rating of 1) applies, then the item is scored a 1. Higher item scores are determined by the number of indicators scored with a Yes response under each of the anchors, 3, 5, and 7 (Exhibit 1).

Under the alternate scoring method, each indicator is individually scored under each of the four anchors, which could extend the assessment time to a total of 4 to 5 hours. This scoring method is often used when the observation focuses on providing detailed feedback to programs or teachers.

Using either scoring method, subscale scores are calculated as the average rating across items for that subscale. The total score is calculated as the average item rating across all items.

Exhibit 1: Item Scoring Based on Indicators for the ECERS-R



Interpretability: Observers must be thoroughly familiar with the ECERS-R, and it is recommended that researchers be trained on the measure and have knowledge of child development and educational implications (Frank Porter Graham Child Development Institute 2005) before interpreting the results. Resources available for interpretation of scores include the Profile and ECERS-R author web site. The Profile graphically displays the scoring information for a comparison of areas of strengths and weaknesses and for the selection of items and subscales targeted for improvement. The Profiles for at least two observations may be plotted side by side to depict changes visually. The manual contains a sample Profile along with a blank Profile and Expanded Score Sheets for photocopying. In addition, the authors maintain an extensive web page (listed under Training Support below) that answers questions about interpretability and use of the scale, and they have published a manual that goes beyond the information available in the instrument document.

Reliability:
(1) Internal consistency: The total scale for internal consistency was 0.92 and ranged from 0.71 (Parents and Staff) to 0.88 (Activities) at the subscale level. The sample of observed classrooms and age intervals were not described.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: Not applicable.
(4) Inter-rater reliability: The average percent agreement over the scale's 470 indicators was 86 percent; no item had an indicator agreement level below 70 percent. At the item level, agreement reached 48 percent exact agreement and 71 percent agreement within 1 point. The correlations for the entire scale were 0.92 (Pearson correlation) and 0.87 (Spearman correlation). Weighted Kappa statistics for individual items ranged from 0.54 to 0.90 for all items except for Using

Language to Develop Reasoning Skills, which had a weighted Kappa of 0.28. Analyses used observation data for 8 to 21 classrooms; age ranges were not given.

Validity Evidence:
The ECERS was revised with data from studies that used the original ECERS, which provided insight into the range of scores on items and the level of item difficulty and validity. Revisions also responded to suggestions from ECERS users.

Construct/Concurrent validity: While the manual presents seven subscales, previous work by the researchers suggests two factors based on 33 items—Teaching and Interactions and Provisions for Learning—in this version and the previous version (Clifford et al. 2005). The ECERS-R manual did not include any other information. (See the Early Childhood Environment Rating Scale-Extension [ECERS-E], Classroom Assessment Scoring System [CLASS], and Teacher Behavior Rating Scale [TBRS] profiles in this compendium for later studies that use the ECERS-R as a comparison measure for validity evidence. See the Caregiver Interaction Scale [CIS] profile in this compendium for subsequent studies that use the ECERS as a comparison measure for validity evidence.)

Predictive validity: The developers note that the previous version (1980) demonstrated predictive validity that is expected to be maintained by the current version (Harms et al. 1998).

Bias Analysis: The ECERS revision process included focus groups of researchers and practitioners that determined how the ECERS functioned in classrooms including children with special needs and culturally diverse children.

Training Support: Observers administering the ECERS-R should be highly trained. Training tools include administration instructions in the manual, training aids from the publisher's web site, and in-person trainings. The Video Observation for the ECERS-R and Instructor's Guide and Video Guide and Training Workbook are available on the publisher's web site. In-person trainings are available during various times of the year. The web site, http://www.fpg.unc.edu/~ecers/, provides information on in-person trainings and links to additional Expanded Score Sheets, Profiles, and other useful information. Observers attending training should have knowledge of child development and educational implications (Frank Porter Graham Child Development Institute 2005). Additionally, researchers should contact the authors regarding separate training to inter-rater reliability with the authors.

Adaptations/Special Instructions for Individuals with Disabilities: Not applicable.

Alternate Forms: Not applicable.

Previous Version: The revised version of the ECERS includes items that are (1) inclusive of children with disabilities and (2) sensitive to cultural diversity. Items added to the new version include the Interaction, Curriculum, Health & Safety, and Parents & Staff scales. Scoring of the ECERS-R is more practical and accurate with the Expanded Score Sheet and additional notes on clarification of scoring (Frank Porter Graham Child Development Institute 2005).

NCEE or REL Study Use:[3] National Evaluation of Early Reading First

[1] Validity evidence refers to the previous version (ECERS; 1980).

[2] The ECERS-R manual did not include information on construct/concurrent validity. The reader may review profiles in the current compendium for the Early Childhood Environment Rating Scale-Extension (ECERS-E), Classroom Assessment Scoring System (CLASS), and Teacher Behavior Rating Scale (TBRS) profiles in this compendium for later studies that use the ECERS-R as a comparison measure for validity evidence. See the Caregiver Interaction Scale (CIS) profile in this compendium for subsequent studies that use the ECERS as a comparison measure for validity evidence.

[3] See Table F.1 for web address.

References:

Clifford, Richard M., Oscar Barbarin, Florence Chang, Diane Early, Donna Bryant, Carollee Howes, Margaret Burchinal, and Robert Pianta. "What Is Pre-Kindergarten? Characteristics of Public Pre-Kindergarten Programs." *Applied Developmental Science,* vol. 9, no. 3, 2005, pp. 131.

Frank Porter Graham Child Development Institute. "Environment Rating Scales." Available at [http://www.fpg.unc.edu/~ecers/]. 2005.

Halle, Tamara, and Jessica Vick. "Quality in Early Childhood Care and Education Settings: A Compendium of Measures." Submitted to the Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Washington, DC: Child Trends, November 2007.

Harms, Thelma, Richard M. Clifford, and Debby Cryer. *Early Childhood Environment Rating Scale, Revised Edition.* New York: Teachers College Press, 1998.

Harms, Thelma, and Debby Cryer. *Early Childhood Environment Rating Scale: Video Guide & Training Workbook.* New York: Teachers College Press, 1999.

Harms, Thelma, and Debby Cryer. *Video Observations for the Early Childhood Environment Rating Scale, Revised Edition.* New York: Teachers College Press, 1999.

Kisker, Ellen E., Kimberly Boller, Charles Nagatoshi, Christine Sciarrino, Vinita Jethwani, Teresa Zavitsky, Melissa Ford, and John M. Love. "Resources for Measuring Services and Outcomes in Head Start Programs Serving Infants and Toddlers." Washington, DC: Mathematica Policy Research, 2003.

Paget, Kathleen D. "Review of ECERS-R." In *The Fourteenth Mental Measurements Yearbook* edited by Barbara S. Plake and James C. Impara. Lincoln, NE: The Buros Institute of Mental Measurements, 2001.

EARLY LANGUAGE & LITERACY CLASSROOM OBSERVATION (ELLCO) PRE-K AND K-3 TOOLS, 2008

Authors: ELLCO Pre-K: Miriam W. Smith, Joanne P. Brady, and Louisa Anastasopoulos ELLCO K-3: Miriam W. Smith, Joanne P. Brady, and Nancy Clark-Chiarelli		**Type of Assessment:** Classroom observation **Domain:** Classroom quality (teacher-student interactions and classroom environment); instructional practice (reading; language arts/language proficiency)
Publisher: Paul H. Brookes Publishing Co. 800-638-3775 http://www.brookespublishing.com		**Grade/Age Range:** Pre-K through grade 3 **Administration Interval:** As frequently as desired
Material, Training, and Scoring Costs: Pre-K Set (5 Observation Tool Booklets and User's Guide): $50.00 K-3 Set (5 Observation Tool Booklets and User's Guide): $50.00		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2 (bachelor's degree with coursework in measurement and domain) **Personnel for Administration:** Highly trained individual **Training for Administration:** Degree or professional experience required (professional with knowledge of children's language and literacy development as well as experience in classroom teaching and conducting classroom observations)
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (<$100) **Time to Administer:** 1 to 1.5 hours **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3[1] (all at or above 0.70) **Predictive Validity:** Available[1] **Construct/Concurrent Validity:** Available[1] **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The ELLCO Pre-K and the ELLCO K-3 are observation instruments designed for assessing the degree to which classroom settings support language and literacy development in pre-kindergarten and the early primary grades. Practitioners and researchers can use the ELLCO Pre-K to observe center-based pre-kindergarten and kindergarten classrooms and the ELLCO K-3 to observe kindergarten through grade 3 classrooms. Both measures consist of a literacy checklist, observation instrument, and teacher interview.

The ELLCO Pre-K observation tool consists of 19 items organized into five sections: (1) Classroom Structure (4 items on classroom organization and contents, children's access to and use of materials, management practices, and adult roles/focus); (2) Curriculum (3 items on curriculum, instructional strategies, child-centeredness, and diversity); (3) Language Environment (4 items on discourse climate, opportunities for extended conversations, vocabulary development, and phonological awareness development); and (4) Books and Book Reading (5 items on organization and use of the book area, characteristics of available books, use of books across curriculum content areas, and frequency and quality of book reading); and (5) Print and Early Writing (3 items on available writing materials, print awareness opportunities, instructional strategies, and use of environmental print).

The ELLCO K-3 observation tool consists of 18 items organized into the five sections described above. The content of items in each section differs slightly to reflect students' advanced developmental status and primary grade classroom practices.

For both measures, each item rates the characteristics of classroom practices along a 5-point scale, ranging from "deficient practice" to "exemplary practice." For each item, observers read anchor statements that describe the practice and the nature and quality of evidence required for a particular rating. They then select from customized descriptive indicators for each scale point. Together, the anchor statements and descriptive indicators provide a rubric for rating item-specific content in a common format across items. Observers may also record notes on Evidence Pages for each item. After completing the observation, observers conduct a brief Teacher Interview to clarify and/or supplement their observations and ratings. Observers typically complete observations in 20 to 45 minutes and the teacher interview in 10 minutes.

Other Languages: None.

Uses of Information: Researchers and evaluators may use the ELLCO Pre-K and the ELLCO K-3 to assess the quality of language and literacy practices in early childhood and early elementary classrooms. The measures also lend themselves to use by supervisors, mentors, and professional development facilitators to assess practices in language and literacy instruction and to inform the planning and implementation of professional development training. Teachers may also use the measures to self-assess their classroom practices.

Methods of Scoring: Scoring procedures are similar for the ELLCO Pre-K and the ELLCO K-3, with both measures scored according to two main subscales: (1) General Classroom Environment (Sections 1 and 2 combined) and (2) Language and Literacy (Sections 3, 4, and 5 combined). Observers rate the individual items using a 5-point scale (1 = deficient practice, 2 = inadequate

practice, 3 = basic practice, 4 = strong practice, and 5 = exemplary practice). Observers complete a score form on which they record subtotals for each section and average scores for the General Classroom Environment subscale and Language and Literacy subscale. Average subscale scores are computed by dividing the total points assigned for the subscale by the number of items in it.

Interpretability: The authors state that an examination of average subscale scores and specific components of the subscale can reveal areas of strength and weakness. These scores may be used to track progress and to plan professional development activities.

Reliability:
Reliability information is not yet available for the ELLCO Pre-K and the ELLCO K-3. The User's Guides for both the ELLCO Pre-K (Smith, Brady, and Anastasopoulos 2008) and the ELLCO K-3 (Smith, Brady, and Clark-Chiarelli 2008) include reliability and validity information for the ELLCO Toolkit, Research Edition (Smith et al. 2002; see Previous Version). The authors conducted reliability studies with data from a sample of 150 preschool classrooms in lower-income communities in New England.
(1) Internal consistency reliability: The authors reported internal consistency estimates for scores based on data collected in 2001 through 2007 with the ELLCO Toolkit, Research Edition (N = 547 to 634 students). Alphas for scores from the Literacy Environment Checklist (Books subtotal, Writing subtotal, and total score) were 0.76, 0.75, and 0.84, respectively. For scores from the Classroom Observation section, the authors reported alphas of 0.84 (General Classroom Environment subtotal score), 0.89 (Language, Literacy, and Curriculum subtotal score), and 0.93 (total score). For scores from the Literacy Activities Rating Scale, alphas were 0.90 (Full-Group Book Reading subtotal score), 0.74 (Writing subtotal score), and 0.72 (total score). The authors reported similar internal consistency coefficients for analyses conducted with data collected from 1997 through 2002 with smaller samples.
(2) Test-retest reliability: The authors cited findings on the stability of comparison group data over time as evidence of test-retest reliability for the Classroom Observation. In a study of the effectiveness of a preschool literacy intervention, comparison group Classroom Observation summary scores remained stable between the fall and spring observations. However, comparison group summary scores did not remain stable over time for the Books subtotal scores and the total scores of the Literacy Environment Checklist or for the Full-Group Book Reading and Writing subtotal scores of the Literacy Activities Rating Scale.
(3) Alternate form reliability: No alternate forms.
(4) Inter-rater reliability: The authors reported estimates of inter-rater reliabilities using percent agreement within one point, finding estimates of 88, 90, and 81 percent, respectively, for ratings on the Literacy Environment Checklist, Classroom Observation, and Literacy Activities Rating Scale of the ELLCO Toolkit, Research Edition.

Validity Evidence:
The ELLCO Pre-K and the ELLCO K-3 User's Guides state that items were designed to capture "important and observable" aspects of language and literacy in preschool and early elementary classrooms. Validity information is not yet available for the ELLCO Pre-K and the ELLCO K-3. The User's Guides for both the ELLCO Pre-K (Smith, Brady and Anastasopoulos 2008) and the ELLCO K-3 (Smith, Brady, and Clark-Chiarelli 2008) cite validity data for the ELLCO Toolkit, Research Edition (Smith et al. 2002; see Previous Version) that were collected from a sample of 150 preschool classrooms in lower-income communities in New England. The guides report

intercorrelations between summary variables of the ELLCO Toolkit, Research Edition. For the Classroom Observation section, its subscale scores for Language, Literacy, and Curriculum and for General Classroom Environment correlated 0.95 and 0.87, respectively, with the total score and 0.69 with each other. For a Literacy Environment Checklist, its Books and Writing subscale scores correlated 0.89 and 0.90, respectively, with the total score and 0.62 with each other. For a Literacy Activities Rating Scale, its subscale scores on Full-Group Book Reading and Writing correlated 0.75 and 0.63 with the total scores.

Construct/Concurrent validity: The authors reported correlations between raw scores on the Learning Environment subscale of the Classroom Profile (Abbott-Shim and Sibley 1998) and the ELLCO Toolkit's General Classroom Environment subtotal scores; Language, Literacy, and Curriculum subtotal scores; and the Classroom Observation Total scores (*r*s = 0.41, 0.31, and 0.44, respectively).

Raw scores on the Scheduling subscale of the Classroom Profile (Abbott-Shim and Sibley 1998) correlated 0.12, 0.09, and 0.07, respectively, with the General Classroom Environment subtotal scores; the Language, Literacy, and Curriculum subtotal scores; and the Classroom Observation total scores of the ELLCO Toolkit, Research Edition. The authors interpreted the low correlations as evidence of divergent validity, recognizing that the ELLCO subscales do not emphasize scheduling over other classroom characteristics (although they do assess it).

The authors also examined change over time in treatment and comparison group summary scores on the Literacy Environment Checklist, Classroom Observation, and Literacy Activities Rating Scale of the ELLCO Toolkit, Research Edition. They compared scores of preschool classrooms participating in a literacy intervention to control group classroom scores. For the Literacy Environment Checklist and the Classroom Observation, intervention classrooms demonstrated significantly higher scores in the spring compared to their own fall scores and compared to the control groups' spring scores. Findings were mixed for the Literacy Activities Rating Scale; the authors reported a significant increase in intervention group scores for the Writing subtotal scores but no significant change in intervention group scores for the Full-Group Book Reading subtotal or for total scores.

Predictive validity: The authors reported results of hierarchical linear modeling analyses examining the contributions of classroom quality (as measured with the Classroom Observation of the ELLCO Toolkit, Research Edition) to Head Start students' receptive vocabulary measured with the Peabody Picture Vocabulary Test—Third Edition and early literacy scores on the Profile of Early Literacy Development. These techniques allowed researchers to distinguish between-classroom variation associated with student background variables (income, gender, age, and at-home language) from between-classroom variation associated with classroom experiences.

The researchers found that 15 percent of the variance in vocabulary scores and 20 percent of the variance in literacy scores was attributable to differences between classrooms. They attributed the variance to classroom factors. Of that variance, the ELLCO Classroom Observation scores accounted for 80 percent of the between-classroom variance in vocabulary and 67 percent of the between-classroom variance in early literacy (Dickinson et al. 2000, as cited in Smith, Brady, and Clark-Chiarelli 2008).

Bias Analysis: No information available.

Training Support: The User's Guides contain guidelines and suggestions for conducting observations, training observers, and establishing inter-rater reliability. Trainees are advised to conduct practice observations in classrooms, preferably with a partner or in a group. The publisher's web site (http://www.brookespublishing.com) advertises optional one- and three-day seminars on use of the ELLCO measures. The seminars are designed for groups of 5 to 20 participants, with speaker fees estimated at $1,500 (one day) and $2,750 to $7,750 (three days). Fees do not include speaker travel costs and may vary with group size (estimates are based on a group of 20) and training activities. In addition, the publisher periodically offers two-day seminars for training of trainers at selected sites across the country. The registration fee for the seminars is $550.

Adaptations/Special Instructions for Individuals with Disabilities: Not applicable.

Alternate Forms: Not applicable.

Previous Version: The ELLCO was first published in 2002 as the ELLCO Toolkit, Research Edition (Smith and Dickinson 2002). It was designed to observe and rate literacy and language instruction in pre-kindergarten through grade 3 classrooms, but most items were geared to pre-kindergarten classrooms. In 2008, the ELLCO Pre-K and the ELLCO K-3 were published as separate instruments. The authors substantially revised the ELLCO K-3 to make it more appropriate for use in kindergarten through grade 3 classrooms (given the substantial revisions, it was published as a research edition). Both the ELLCO Pre-K and the ELLCO K-3 have incorporated the Literacy Environment Checklist and Literacy Activities Rating Scale into the observation protocol. In addition, the current version includes detailed descriptors for all five scale points rather than for three scale points. The authors state that, compared to the original version, the ELLCO Pre-K items place greater emphasis on phonological awareness, efforts to increase spoken vocabulary, and uses of environmental print, noting that the ELLCO K-3 items better assess evidence-based approaches to reading and writing instruction.

NCEE or REL Study Use:[2] The Effects of Opening the World of Learning (OWL) on the Early Literacy Skills of At-Risk Urban Preschool Students

[1] Reliability and validity studies were conducted with the ELLCO Toolkit, Research Edition (Smith et al. 2002; see Previous Version). Reliability and validity information is not yet available for the ELLCO Pre-K and the ELLCO K-3

[2] See Table F.1 for web address.

References:

Dickinson, David K., Kimberley Sprague, Aline Sayer, Candy Miller, N. Clark, and A. Wolf. "Classroom Factors that Foster Literacy and Social Development of Children from Different Language Backgrounds." In Marita Hopmann (chair), *Dimensions of Program Quality that Foster Child Development: Reports from 5 Years of the Head*

Start Quality Research Centers. Poster session presented at the biennial National Head Start Research Conference, Washington, DC, 2000.

Smith, Miriam W., Joanne P. Brady, and Louisa Anastasopoulos. *Early Language & Literacy Classroom Observation Pre-K Tool*. Baltimore: Paul H. Brookes Publishing Co., 2008.

Smith, Miriam W., Joanne P. Brady, and Louisa Anastasopoulos. *User's Guide to the Early Language & Literacy Classroom Observation Pre-K Tool*. Baltimore: Paul H. Brookes Publishing Co., 2008.

Smith, Miriam W., Joanne P. Brady, and Nancy Clark-Chiarelli. *Early Language & Literacy Classroom Observation K-3 Tool (Research Edition)*. Baltimore: Paul H. Brookes Publishing Co., 2008.

Smith, Miriam W., Joanne P. Brady, and Nancy Clark-Chiarelli. *User's Guide to the Early Language & Literacy Classroom Observation K-3 Tool (Research Edition)*. Baltimore: Paul H. Brookes Publishing Co., 2008.

Smith, Miriam W., and David K. Dickinson (with A. Sangeorge and Louisa Anastasopoulos). *Early Language and Literacy Classroom Observation (ELLCO) Toolkit (Research Ed.)*. Baltimore: Paul H. Brookes Publishing Co., 2002.

EARLY READING PROFESSIONAL DEVELOPMENT (PD) CLASSROOM OBSERVATION, 2008

Authors: Michael S. Garet, Stephanie Cronen, Marian Eaton, Anja Kurki, Meredith Ludwig, Wehmah Jones, Kazuaki Uekawa, Audrey Falk, Howard Bloom, Fred Doolittle, Pei Zhu, and Laura Sztejnberg		**Type of Assessment:** Classroom observation **Domain:** Instructional practices (reading)
Publisher: Described in Garet et al. 2008.		**Grade/Age Range:** Used by authors in grade 2 classrooms **Administration Interval:** As frequently as desired
Material, Training, and Scoring Costs: Not specified		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2 (bachelor's degree with coursework in measurement and domain) **Personnel for Administration:** Highly trained individual **Training for Administration:** Degree or professional experience required (bachelor's degree or higher and graduate or professional research training)
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** To be determined upon negotiation with the publisher **Time to Administer:** An entire day's reading instruction **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Early Reading Professional Development (PD) Interventions Study classroom observation protocol assesses teachers' classroom implementation of the instructional practices and content in the study's professional development curriculum. The protocol's content is based on recommended instructional practices for teachers of early readers (Moats 2005), empirical knowledge of reading research (Armbruster et al. 2001), and specific strategies presented in the study's professional development curriculum. The protocol assesses to what degree, and through what instructional practices, teachers focus instruction on the five foundational reading skills identified by the National Reading Panel that underlie the study's professional development curriculum: (1) phonemic awareness, (2) phonics, (3) fluency, (4) comprehension, and (5) vocabulary. It also assesses levels of student engagement during observed lessons and whether teachers differentiate instruction for different types of learners. Garet et al. (2008) developed and used the protocol to conduct low-inference observations of grade 2 teachers' instructional practices during reading lessons. The protocol is used during observations of a class's entire reading instruction period during one school day. The protocol reserves space to record up to 60 3-minute intervals, or 180 minutes of instruction). In the Early Reading PD Interventions Study, the length of classroom observations ranged from 16 to 60 3-minute intervals (i.e., 48 to 180 minutes), with a mean of 39.4 intervals (i.e., 118 minutes).

The protocol consists of four parts, but only Part II was used for the construction of scales in the current study. Part I is a checklist to be completed before the observed reading lesson. It documents details about the lesson to be observed, including materials used, grouping of students, and potential participation of support personnel during instruction. Part II is to be completed during the lesson. During 3-minute intervals, the observer marks the components of reading instruction on which the lesson focuses (phonemic awareness, phonics, fluency, reading comprehension, vocabulary, other instruction); whether the teacher used component-specific instructional practices; whether the teacher differentiated instruction; the instructional format (whole class, small groups, pairs, teacher working with particular student(s), break-in instruction); instructional materials; and the number of students observed to be off-task during the observation. Part III is a checklist of reading program implementation. In Part IV, observers record subjective opinions of the observed instruction.

The authors used data from Part II (interval samples) to create three scales: Explicit Instruction, Independent Student Activity, and Differentiated Instruction. The Explicit Instruction scale consists of 27 items (4 phonemic awareness items, 5 phonics items, 3 fluency items, 11 comprehension items, 4 vocabulary items). Examples of items on the scale include directly explaining phonics patterns and how story events support predictions of what will happen next. The Independent Student Activity scale consists of 25 items describing what the student is doing in relation to the content (2 phonemic awareness items, 3 phonics items, 6 fluency items, 11 comprehension items, 3 vocabulary items). Examples of items on the scale include practicing decoding independently, completing a graphic organizer, or reading a passage. The Differentiated Instruction scale consists of two items (i.e., differentiated materials are used, teacher works separately with a particular student or group of students), both of which must be recorded for an interval to represent differentiated instruction.

Other Languages: None.

Uses of Information: The authors have used the protocol to assess teachers' classroom implementation of the practices taught in the study's professional development curriculum. In addition to computing scale scores for Explicit Instruction, Independent Student Activity, and Differentiated Instruction, the authors calculated descriptive statistics for different aspects of the observations of classroom reading lessons, including frequency of time spent in different classroom formats (whole group, small groups, and so forth) and components of reading instruction (phonemic awareness, phonics, fluency, reading comprehension, vocabulary, other instruction).

Methods of Scoring: For each three-minute interval, observers note the presence or absence of the 27 Explicit Instruction practices, 25 Independent Student Activities, and/or 2 Differentiated Instruction practices. When constructing the Explicit Instruction and Independent Student Activity scales, the authors included only intervals that focused on one of the five components of reading instruction (phonemic awareness, phonics, fluency, comprehension, vocabulary) and excluded intervals covering other language arts subjects or non-instructional activities. They included all observed intervals for the Differentiated Instruction scale. They derived scale scores by using a Rasch model at level one of a two-level model. Due to a potential confound of content (reading instruction component) and instructional approaches (explicit instruction or independent student activity), the authors estimated models for the two instructional approaches by using the frequency with which the teachers engaged in practices identified with these instructional approaches, controlling for the reading instruction component. That is, the "teacher's log odds of engaging in [explicit instruction or independent student activity] during a 3-minute interval is modeled as a function of the reading instruction component. . .and the teacher's latent propensity to engage in [explicit instruction or independent student activity]" (Garet et al. 2008, p. F-1). Given that most teachers did not engage in differentiated instruction during any observed intervals, the authors created scale scores by ". . .computing a percentage of intervals in which differentiated instruction took place, adjusting for the relative prevalence of differentiated instruction across the sample in the particular components (for example, phonemic awareness, comprehension) in which the teacher provided instruction" (p. F-3).

Interpretability: A teacher's scale score on the Explicit Instruction scale represents the teacher's predicted log odds of engaging in explicit instruction during an interval, controlling for the component of instruction. The authors used a similar model, substituting independent student activity for explicit instruction, to derive a scale score for that scale. The Differentiated Instruction scale score represents the teacher's use of differentiated instruction relative to the use by all teachers in a particular content area; that is, it is a weighted average of the proportion of intervals in which differentiated instruction was observed (adjusted for the relative frequency of differentiated instruction in specific components). In all three scales, a higher score means that the teacher is more likely to implement the practice.

Reliability:
(1) Internal consistency reliability: For the Explicit Instruction and Independent Student Activity scales, the reliability estimates for Explicit Instruction scale raw scores were 0.83 (fall 2005), 0.80 (spring 2006), and 0.78 (fall 2006). For the Independent Student Activity scale, reliability estimates for raw scores were 0.81 (fall 2005), 0.74 (spring 2006), and 0.72 (fall 2006). Most teachers in the study did not differentiate instruction during observed lessons; therefore, the authors estimated the reliability of raw scores from the Differentiated Instruction scale by using

data from teachers who were observed to differentiate instruction and who were included in the outcomes analysis samples (N = 253 in fall 2005, N = 248 in spring 2006, and N = 228 in fall 2006; see Validity Evidence for information about the full sample). Reliability estimates for raw scores from the Differentiated Instruction scale were 0.88 (fall 2005), 0.89 (spring 2006), and 0.90 (fall 2006).
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: Not applicable.
(4) Inter-rater reliability: Both a gold standard observer and a regular observer coded 10 percent of the 730 study observations. The authors analyzed the data for percent agreement. With the observations coded by three-minute intervals and the availability of several potential codes for each interval, the authors observed a large number of empty protocol cells. Noting that any inter-rater calculation that equally weights coded and empty cells would overestimate the degree of agreement, the authors calculated percent agreement between observers on cells in which one or both observers coded that instruction occurred during the interval. For data collected in fall 2005 (N = 25 pairs of observations), spring 2006 (N = 26 pairs of observations), and fall 2006 (N = 22 pairs of observations), percentage agreement for the overall observation protocol ranged from 90.4 to 91.0.

Validity Evidence:
The authors developed the protocol items to reflect the major components and characteristics of the study's professional development curriculum. The curriculum and observation protocol were based on Moats' (2005) training series and current research on teaching reading to beginning readers, with emphasis on the five foundational reading skills identified by the National Reading Panel (phonemic awareness, phonics, fluency, comprehension, vocabulary). Observers used the protocol to conduct observations for the Early Reading PD Interventions study in grade 2 classrooms in 90 schools in six school districts. The districts were located in urban or urban fringe areas in four Eastern and Midwestern states and served substantial populations of English-proficient students from low-income households. Ninety-three percent of the schools were Title I schools, and 78 percent of the students were eligible for free or reduced-price lunch. Most students in sample schools were Black (78.4 percent); with 15.6 percent White. Observations were conducted in regular (non-special education) grade 2 classrooms, with sample sizes ranging from 250 to 270 teachers throughout the study.

Construct/Concurrent validity: While no information was provided comparing the current measure's scores other measures, Garet et al. (2008) compared observational data collected with the protocol from teachers in treatment and control groups at the end of the school year. Treatment group teachers received professional development training alone (treatment A) or training and coaching (treatment B). Control group teachers had no exposure to the study's professional development training. In their first analysis, Garet et al. used a two-level hierarchical model with teachers nested within schools to determine possible impacts of the professional development intervention on observed teacher instructional practices during the intervention year. They found that teachers who received treatment A engaged in significantly more explicit instruction than did control group teachers (51 versus 42 percent of observed intervals, respectively). Teachers who received treatment B also engaged in significantly more explicit instruction than did control group teachers (57 versus 42 percent of observed intervals, respectively). The authors reported no significant effects for treatment A or B on the use of independent student activity or differentiated instruction in the classroom. An identical followup

analysis with observational data collected in the fall of the year following the training intervention showed no significant effects.

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: For the current study, observers included graduate students at one site and then study personnel with at least a bachelor's degree or higher and research training. Garet et al. (2008) described training workshops that they conducted for observers. Study lead observers, who subsequently served as the gold standard, participated in 10 days of training workshops (6 days of reading instruction content and use of the protocol and 4 days of practice in classrooms). Study team observers participated in 5 days of training, 2 of which were dedicated to classroom practice. The training reviewed the five components of reading instruction included in the protocol as well as observers' roles and responsibilities in the study. It also included demonstrations and discussion of coding strategies and opportunities to practice coding in classrooms and with videotapes of classroom reading lessons. Training materials included a training manual, PowerPoint presentations, and handouts. The researchers conducted followup training six months and one year after the original training. The followup training addressed issues that arose during the first wave of observations and reviewed and reinforced coding procedures.

Adaptations/Special Instructions for Individuals with Disabilities: Not applicable.

Alternate Forms: Not applicable.

Previous Version: None.

NCEE or REL Study Use:[1] Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement

[1] See Table F.1 for web address.

References:

Armbruster, Bonnie B., Fran Lehr, and Jean Osborn. "Put Reading First: The Research Building Blocks for Teaching Children to Read, K-3." Washington, DC: The Partnership for Reading: National Institute for Literacy; National Institute of Child Health and Human Development; and U.S. Department of Education, 2001.

Garet, Michael S., Stephanie Cronen, Marian Eaton, Anja Kurki, Meredith Ludwig, Wehmah Jones, Kazuaki Uekawa, Audrey Falk, Howard Bloom, Fred Doolittle, Pei Zhu, and Laura Sztejnberg. "The Impact of Two Professional Development Interventions of Early Reading Instruction and Achievement." (NCEE 2008-4030). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2008.

Moats, Louisa C. “Language Essentials for Teachers of Reading and Spelling (LETRS).” Longmont, CO: Sopris West Educational Services, 2005.

INFANT/TODDLER ENVIRONMENT RATING SCALE
REVISED EDITION, 2006

Authors: Thelma Harms, Debby Cryer, and Richard M. Clifford		**Type of Assessment:** Classroom observation, with some teacher report **Domain:** Classroom quality (classroom environment with some teacher-student interaction)
Publisher: Teachers College Press 800-575-6566 http://www.teacherscollegepress.com		**Grade/Age Range:** Birth to 30 months **Administration Interval:** As frequently as desired
Material, Training, and Scoring Costs: ITERS-R rating scale in spiral binder with Expanded Score Sheet and Profile for photocopying: $19.95 Video Observations for the ITERS-R (DVD/VHS) and Instructor's Guide: $59 Video Guide and Training Workbook: $4 ITERS-R training (excluding travel) at the University of North Carolina (UNC): From $825 to $1,225 depending on focus of the training		**Personnel and Training Requirements** **Credentials Required for Use:** Level 1 (training or supervised experience with measurement) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours) The authors recommend that observers attend a training session (with one or more practice observations) led by an experienced ITERS-R trainer. Researchers should contact the authors regarding training to evaluate inter-rater reliability. In addition, observers attending training should have knowledge of child development and educational implications (Frank Porter Graham Child Development Institute 2005).
Languages: English, German, Japanese, and Spanish		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (under $100) **Time to Administer:** 2 to 5 hours depending on scoring option **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3[1] (all at or above 0.70) **Predictive Validity:** Available[2] **Construct/Concurrent Validity:** Available[2] **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The ITERS-R is a classroom assessment tool designed to measure the quality of group programs for infants and toddlers (birth to 30 months) by collecting data through classroom observation and a staff interview. The assessment is a 39-item rating scale organized into seven environmental subscales: (1) Space and Furnishings, (2) Personal Care Routines, (3) Listening and Talking, (4) Activities, (5) Interaction, (6) Program Structure, and (7) Parents and Staff. Each item has several quality indicators, accounting for a total 467 Yes/No indicators. Administration time varies with the scoring option and whether the assessor is an outside observer, although average administration time is 3.5 hours, including the staff interview. Reviewers caution against practitioners' use of the assessment because it does not describe the validity of the measure in detail and instead relies on the validity research of the original version (Carey 2007; Kush 2007).

Other Languages: Although the ITERS-R was translated into other languages (German, Japanese, and Spanish), there is no information regarding the development of the scales in those languages or investigations of comparability of scores with the English version.

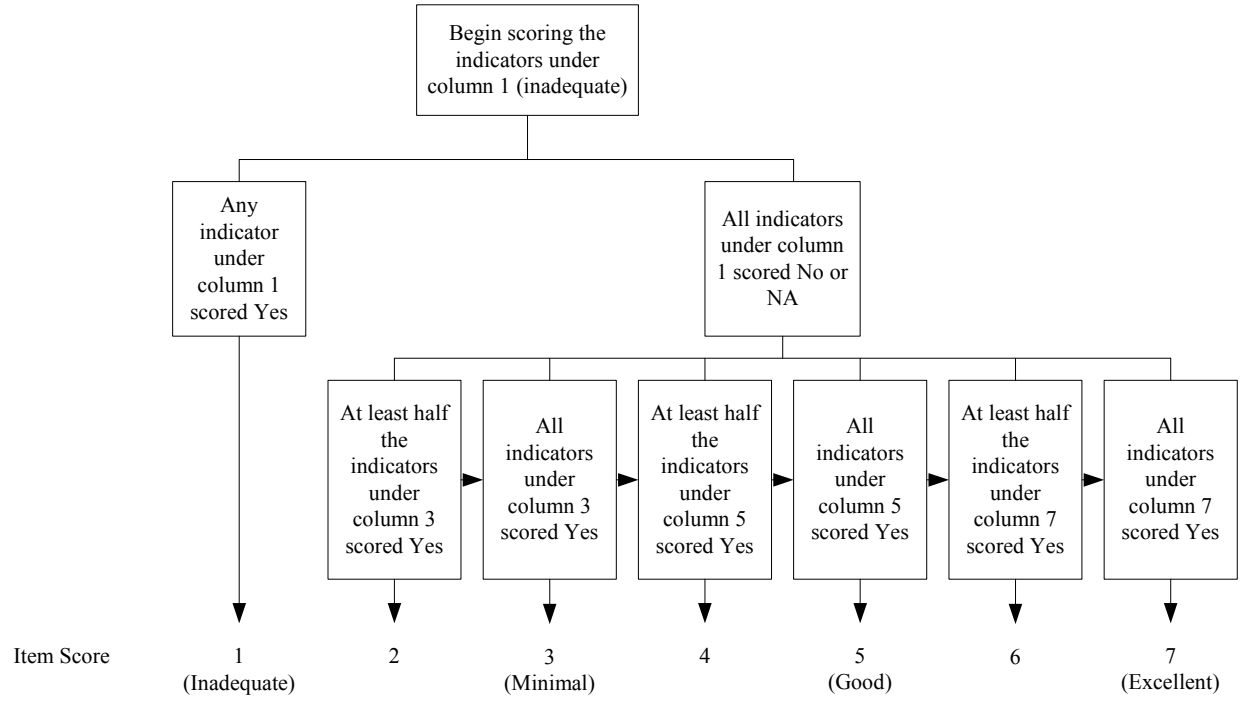
Uses of Information: The assessment may be used by program staff as a self-assessment tool and by outside observers for program monitoring, evaluation, development, and research.

Methods of Scoring: An individual thoroughly familiar with the ITERS-R should score the assessment. The Expanded Score Sheet is used to record the ratings for quality indicators, items, subscale scores, and total scores as well as any observer comments. The indicators, which have Yes/No/Not Applicable (NA) response choices, are used to score the items from 1 (Inadequate) to 7 (Excellent). Indicators fall under columns at the scale anchors 1, 3, 5, and 7. Items may be scored two ways as described in detail in the manual. Under the standard scoring option for each item, if any of the indicators in the Inadequate column (or rating of 1) applies, then the item is scored a 1. Higher item scores are determined by the number of indicators scored with a Yes response under each of the anchors, 3, 5, and 7 (Exhibit 1).

Under the alternate scoring method, each indicator is individually scored under each of the four anchors, which could extend the assessment time to a total of 4 to 5 hours. This scoring method is often used when the observation focuses on providing detailed feedback to programs or teachers.

Using either scoring method, subscale scores are calculated as the average rating across items for that subscale. The total score is calculated as the average item rating across all items.

Exhibit 1: Item Scoring Based on Indicators for the ITERS-R



Interpretability: Resources available for interpretation of scores include the Profile and the authors' web site. Observers must be thoroughly familiar with the ITERS-R. In addition, they are advised to be trained on the measure and demonstrate knowledge of child development and educational implications (Frank Porter Graham Child Development Institute 2005) before interpreting the results. The Profile graphically displays the scoring information to permit a comparison of areas of strengths and weaknesses and the selection of items and subscales for improvement. The Profiles for at least two observations may be plotted side by side for visual depiction. A sample Profile appears in the manual along with blank Profile and Expanded Score Sheets for photocopying. In addition, the authors maintain an extensive web page (listed under Training Support) that answers questions about interpretability and use of the scale, and they have published a manual that goes beyond the information available in the instrument document.

Reliability: Reliability studies apply to the ITERS-R (2003). The items and indicators for the 2006 version of the ITERS-R are the same as for the 2003 version (see Previous Version for more information).
(1) Internal consistency reliability: The total scale internal consistency was 0.93, and the internal consistency for the child-related items (items 1 through 32) was 0.92. Subscale internal consistency reliability ranged from 0.47 (Space and Furnishings) to 0.80 (Interaction), with four of the seven subscales at or above 0.70.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: Not applicable.
(4) Inter-rater reliability: A two-phase pilot study conducted in 2001 and 2002 estimated the level of data reliability from the ITERS-R. In the first phase, 10 trained observers conducted 12 observations (in groups of 2 or 3) in 9 infant and/or toddler centers using the first version of the ITERS-R. The authors describe the second phase as more formal; 6 trained observers conducted

45 paired observations, each lasting about 3.5 hours (including teacher report). The reliability analysis involved 34 centers, of which 7 included children identified with disabilities. Reliability analyses were based on 90 observations using the 6 trained observers from phase two of the pilot study. Authors calculated interclass correlations, percentage agreement, and weighted Kappa statistics for inter-rater reliability. The interclass correlation was 0.92 for ratings based on the full scale as well as for the child-related items. Interclass correlations for ratings by subscale scores ranged from 0.67 (Personal Care Routines) to 0.92 (Parents and Staff). Authors also calculated percentage agreement within one point for paired observations. There was agreement within 1 point 85 percent of the time across the full scale and 83 percent of the time across the 32 child-related items. Item agreement within 1 point ranged from 64 percent (Item 4: Room arrangement) to 98 percent (Item 38: Evaluation of staff). The weighted Kappa statistic for the full scale was 0.58 and 0.55 for the child-related scale. Two of the weighted Kappa statistics were below 0.40 (0.14 for Item 9: Diapering/toileting and 0.20 for Item 11: Safety practices). Item 34 (Provisions for personal needs of staff) had the highest weighted Kappa statistics at 0.92. All Yes/No indicators achieved agreement 91.7 percent of the time, and child-related indicators achieved agreement 90.3 percent of the time. Item 11 (Safety practices) was the only item with indicator agreement less than 80 percent of the time (79.1 percent); and Item 35 (Staff professional needs) had the highest indicator agreement at 97.4 percent of the time.

Validity Evidence:
Validity studies apply to the original version of the ITERS (1990; see Previous Version for more information). To aid in classifying and assessing quality, revision of the ITERS was based on research evidence from several relevant fields (e.g., health and education), best practices from professionals, and practical constraints of real life in child care settings. The revision process used four sources: (1) research on development in the early years and results associated with the impact of child care environments on children's health and development; (2) a comparison of the content of the original ITERS and assessments by using a similar age group along with documents describing program quality; (3) feedback via web site questionnaires and focus groups of professionals familiar with the ITERS; and (4) use of the ITERS over a two-year period by the co-authors and over 25 trained assessors on the ITERS for the North Carolina Rated License Project. Research and development provided information on the range of scores for certain items in addition to items' level of difficulty and validity. The content comparison identified items to be added or eliminated. Revision of the original scale was based on results from the first phase of the reliability pilot study (see Inter-rater reliability); results from the second phase of the study resulted in the improvement of items with weighted Kappa statistics below 0.50 in order to improve reliability. The printed version of the scale specifies these changes.

Construct/Concurrent validity: The authors state that concurrent validity was established with the original version of the ITERS. Given the ITERS-R's similarity to the ITERS, studies on the ITERS-R have focused on the degree to which trained observers may continue to use the scale reliably (Harms et al. 2003). No information was provided regarding convergent or divergent validity.

Predictive validity: The authors state that predictive validity was established with the original version of the ITERS. Given the ITERS-R's similarity to the ITERS, studies on the ITERS-R

have focused on the degree to which trained observers may continue to use the scale reliably (Harms et al. 2003).

Bias Analysis: The focus groups (mentioned under Content validity) were included in the revision process to determine how the ITERS-R functioned in classrooms including children with special needs and culturally diverse children.

Training Support: Individuals administering the ITERS-R should be highly trained. Training tools for the ITERS-R include the administration instructions presented in the manual, training aids from the publisher's web site, and in-person trainings. The Video Observations for ITERS-R, Instructor's Guide, Video Guide, and Training Workbook are available on the publisher's web site. The Video Observations for the ITERS-R DVD/VHS and the Instructor's Guide demonstrate how to present training activities and answer frequently asked questions about the ITERS-R. In-person trainings are available during various times of the year. The authors' web site (http://www.fpg.unc.edu/~ecers/) provides information on in-person trainings and links to additional Expanded Score Sheets, Profiles, and other useful information. Observers participating in training sessions should demonstrate knowledge of child development and educational implications (Frank Porter Graham Child Development Institute 2005). In addition, researchers should contact the authors about training to evaluate inter-rater reliability with the authors.

Adaptations/Special Instructions for Individuals with Disabilities: Not applicable.

Alternate Forms: Not applicable.

Previous Version: Several differences exist between the original ITERS (1990) and the revised versions. The ITERS-R (2003) updated the scoring for indicators to reflect observed strengths and weaknesses within items; removed negative indicators from all levels except for level 1 (Inadequate); lengthened the Notes for Clarification to enhance clarification; included culturally sensitive items and examples; added new items to some subscales (Listening and Talking, Activities, Program Structure, and Parents and Staff subscales); combined items in the Space and Furnishings subscales in instances of apparent overlap; dropped two items from the Personal Care Routines subscale; and made more gradual the scaling of some items in the Personal Care Routines subscale. In addition, items appear on separate pages followed by the Notes for Clarification, and sample interview questions are included for difficult-to-observe indicators. The ITERS-R (2006) features a new spiral binding, additional Notes for Clarification, and an Expanded Score Sheet, which includes notes and tables to assist with scoring. The items and indicators remain the same as the 2003 version of the ITERS-R.

NCEE or REL Study Use:[3] Evaluating the Impact of the Program for Infant/Toddler Care

[1] The rating refers to the reliability for the total test scores or scores commonly reported. The individual subscales encompassed some ratings below the 0.70 level.

[2] Validity studies apply to the original version of the ITERS (1990; see Previous Version).

[3] See Table F.1 for web address.

References:

Carey, Karen. "Review of ITERS-R." In *The Seventeenth Mental Measurements Yearbook*, edited by Kurt F. Geisinger, Robert A. Spies, Janet F. Carlson, and Barbara S. Plake. Lincoln, NE: The Buros Institute of Mental Measurements, 2007.

Frank Porter Graham Child Development Institute. "Environment Rating Scales." Available at [http://www.fpg.unc.edu/~ecers/]. 2005.

Halle, Tamara, and Jessica Vick. "Quality in Early Childhood Care and Education Settings: A Compendium of Measures." Submitted to the Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Washington, DC: Child Trends, November 2007.

Harms, Thelma, and Debby Cryer. *Infant/Toddler Environment Rating Scale: Video Guide & Training Workbook*. New York: Teachers College Press, 2003.

Harms, Thelma, and Debby Cryer. *Video Observations for the Infant/Toddler Environment Rating Scale, Revised Edition*. New York: Teachers College Press, 2006.

Harms, Thelma, Debby Cryer, and Richard M. Clifford. *Infant/Toddler Environment Rating Scale, Revised Edition*. New York: Teachers College Press, 2003.

Harms, Thelma, Debby Cryer, and Richard M. Clifford. *Infant/Toddler Environment Rating Scale, Revised Edition*. New York: Teachers College Press, 2006.

Kisker, Ellen E., Kimberly Boller, Charles Nagatoshi, Christine Sciarrino, Vinita Jethwani, Teresa Zavitsky, Melissa Ford, and John M. Love. "Resources for Measuring Services and Outcomes in Head Start Programs Serving Infants and Toddlers." Washington, DC: Mathematica Policy Research, 2003.

Kush, Joseph C. "Review of ITERS-R." In *The Seventeenth Mental Measurements Yearbook*, edited by Kurt F. Geisinger, Robert A. Spies, Janet F. Carlson, and Barbara S. Plake. Lincoln, NE: The Buros Institute of Mental Measurements, 2007.

LITERACY OBSERVATION TOOL (LOT)—E-LOT, LOT, AND A-LOT

Authors: E-LOT: Anna W. Grehan, Kimberly J. Motschman, and Lana J. Smith LOT: Anna W. Grehan, Steven M. Ross, and Lana J. Smith A-LOT: Anna W. Grehan, Lisa Dyson, and Lana J. Smith		**Type of Assessment:** Classroom observation **Domain:** Instructional practices (reading, language arts/language proficiency), classroom quality (teacher-student interactions, classroom environment), school engagement
Publisher: Center for Research in Educational Policy (CREP) University of Memphis 866-670-6147 http://www.memphis.edu/crep/		**Grade/Age Range:** Preschool (E-LOT), elementary school (LOT), and middle and high school (A-LOT) **Administration Interval:** As frequently as desired
Material, Training, and Scoring Costs:[1] Training manual/observation forms: $20 per person Training session (excluding travel): $2,000 per 40 participants Observations by CREP: $300 per local observation; additional costs for non-local observations tailored to project Processing fee: $350 per school report if data are entered using online system; $450 if data are entered using paper scantron forms		**Personnel and Training Requirements** **Credentials Required for Use:** Level 1 (training or supervised experience with measurement) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours) E-LOT training includes reading the manual, attending a formal training session, conducting practice observations, and obtaining inter-rater reliability consensus with another observer (Grehan et al. 2006).
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 4 (>$500) **Time to Administer:** 10 to 15 minutes per classroom, with 7 to 10 classrooms, for a total of 3 hours (see Description) **Ease of Administration and Scoring:** 4 (administered or scored by a specialist) **Reliability:** 3[2] (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Literacy Observation Tool (LOT) consists of three instruments—the E-LOT (preschool classrooms), LOT (elementary classrooms), and A-LOT (middle and high school classrooms)—that are used to observe pre-literacy and literacy teacher practices. For the E-LOT and LOT, observers conduct 10-minute observations of 7 to 9 different classrooms during learning center time and literacy teaching blocks. For the A-LOT, observers conduct 15-minute observations of 8 to 10 different classrooms within three-hour blocks of time; developers recommend that one-third of the observations should occur during content-area subjects. Observers complete two forms to assess the extent to which subscale components (described below) are present during the observation period: (1) an individual Classroom Observation Notes Form for each classroom observation and (2) a Data Summary Form, which synthesizes findings from the Classroom Observation Notes Forms. On each Classroom Observation Notes Form, observers write an "O" if the component was observed and comment on the extent to which the teacher emphasized the component. Exceptions apply for one component on each form in which the LOT and A-LOT Materials Used component has a check-all-that-apply response format and the E-LOT Types of Centers component responses include "not in use," "in use," and "explicit connection to literacy skills" instead of an "O" option. Observers then qualitatively synthesize these recorded observations and notes and use a rubric to rate the degree to which they observed each component across all classrooms: 0 = not observed, 1 = rarely, 2 = occasionally, 3 = frequently, and 4 = extensively observed. The manual describes two prominent factors for rating each LOT item: the number of classrooms in which a component was observed and the time and emphasis given to the component within classes. Observers record ratings on the Data Summary Form and may add qualitative comments such as strengths, concerns, impressions about children's progress, and recommendations.

The manuals provide detail on the various E-LOT, LOT, and A-LOT subscales (which developers refer to as observation "categories"), components (which developers refer to as classroom "strategies" or "events"), and individual items. In particular:

The **E-LOT** includes 82 items across four subscales: (1) Instructional Orientation, (2) Instructional Components, (3) Learning Centers, and (4) Reflections. Instructional Orientation (4 items) refers to designated spaces with a classroom containing materials and resources that support language and literacy development (e.g., whole class or small group). The Instructional Components (29 items) contribute to effective and developmentally appropriate literacy development and include Concepts of Print, Alphabetic and Phonological Awareness, Fluency, Vocabulary and Oral Language Development, Development of Cognition and Text Comprehension, and Emergent Writing. The Learning Centers components (41 items) include Other Personnel Assisting with Centers, Center Structure (e.g., student-selected or teacher-assigned), Types of Centers (e.g., art or music), Teacher Interactions during Center Time (e.g., guidance or monitoring), and Student Activities (e.g., engages in small group discussion or listens to stories). The Reflections components (8 items) include Classroom Environment/Climate and Visible Print Environment.

The **LOT** includes 41 items across six subscales, including (1) Instructional Orientation (4 items, e.g., whole class or small group), (2) Instructional Components (14 items on five components for Concepts of Print, Alphabetics, Fluency, Vocabulary, and Text Comprehension), (3) Student

Activities (3 items, e.g., reads self-selected materials or writes independently in response to reading), (4) Learning Environment (4 items, e.g., students actively engaged or teacher actively monitors), (5) Visible Print Environment (4 items), and (6) Materials Used (12 items).

The **A-LOT** includes 58 items across eight subscales: (1) Instructional Orientation (4 items), (2) Instructional Strategies (13 items for seven strategies Fluency, Vocabulary, Text Comprehension, Explicit Content-Area Literacy Instruction, Integration of Subject Areas, Teacher Acts as Coach/Facilitator, and Higher-Order Instructional Feedback), (3) Student Activities (8 items, e.g., sustained reading or experiential hands-on learning), (4) Assessment (2 items), (5) Explicit Writing Instruction (4 items), (6) Student Writing Activities (4 items), (7) Summary of Learning Environment (6 items, e.g., effective classroom management or students actively engaged), and (8) Materials Used (17 items).

Other Languages: None.

Uses of Information: The LOT is an observation instrument that measures instructional practices, student activities, and environmental settings in classrooms where teachers focus on reading and literacy processes. Classroom data are synthesized at the school level so that schools may evaluate the effectiveness of teacher implementation of research-based reading strategies (Grehan and Sterbinsky 2005).

Methods of Scoring: Observers record ratings on the Data Summary Form scantron sheets, attach their individual Classroom Observations Notes Forms, and submit the forms to the developers at the Center for Research in Educational Policy (CREP), College of Education, University of Memphis for scanning, scoring, and analysis. Alternatively, observers may enter Data Summary Form ratings by using an online system. Developers aggregate Data Summary Form results at the school level and report, for each item, the percentage of Data Summary Forms with each type of observation rating (ranging from 0 = not observed to 4 = extensively observed). For example, if observers submitted three Data Summary Forms from a school and one Data Summary Form indicated that an item was not observed and two Data Summary Forms indicated that an item was extensively observed, then the score for that item would be 33.3 percent not observed and 66.6 percent extensively observed.

Interpretability: Developers provide each school with one annual school-level report. Reports are user-friendly documents with tables organized by subscale, component, and item. They show items with a high prevalence of extensively and frequently observed ratings and items with a high prevalence of rarely and not observed ratings. Developers also summarize qualitative comments from observers, if available; at an additional cost, they provide reports of change over time (e.g., spring to fall), custom narrative analysis, and recommendations.

Reliability:
(1) Internal consistency: No information available.
(2) Test re-test reliability: Grehan and Sterbinsky (2005) estimated the reliability for LOT ratings in a sample of prekindergarten through grade 3 classrooms in Tennessee during the 2002–2003 school year. Using the Generalizability Theory framework, the authors calculated a phi coefficient of 0.75 based on five LOTs (two observations in fall and three in spring). They then

extrapolated reliability estimates for different numbers of LOTs; phi coefficients for 1, 3, 6, 8, 10, and 20 LOTs were 0.39, 0.65, 0.78, 0.82, 0.85, and 0.92, respectively.
(3) Alternate forms: Not applicable.
(4) Inter-rater reliability: Huang et al. (2007) estimated inter-rater reliability for ratings of the E-LOT Data Summary Form among 15 pairs of observers at 15 schools in a large urban school district. Overall, the authors used item ratings from 0 (not observed) to 4 (extensively), except for the Types of Centers component, which were coded as 0 (in use) or 1 (not in use). For each E-LOT component, the authors calculated intraclass coefficients based on average ratings across 15 pairs of observers. Most coefficients were greater than 0.70, and 10 items were excluded because of no variance across schools or observers. Coefficients ranged from 0.27 to 1.00, with the exception of one item with a coefficient of -0.11, "reviews vocabulary including environmental print and word walls" in the Vocabulary and Oral Language Development component. The authors explained that the negative coefficient resulted from greater within-school variance than between-school variance, perhaps indicating that observers experienced problems with item term definitions. The authors also presented Kappa and unweighted Kappa statistics and percentage agreement statistics for each of the 15 pairs of observers. Weighted Kappa statistics ranged from 0.68 to 0.98, unweighted Kappa statistics from 0.66 to 0.97, and percentage agreement from 76 to 98 percent.

Validity Evidence:
CREP researchers developed the LOT and worked with researchers and practitioners from Memphis city schools and the Tennessee, Louisiana, and Illinois Departments of Education to assess the measure's content validity (Halle and Vick 2007). LOT and E-LOT Instructional Components are aligned with scientifically-based topic areas identified by the National Reading Panel, National Research Council, and Reading First and Early Reading First. The LOT has since been used as a classroom observation tool in several studies (Grehan et al. 2006; Grehan and Ross 2004; and Grehan et al. 2007).

Construct/Concurrent validity: Descriptive studies suggest that E-LOT scores are positively correlated with student achievement and converge with the classroom observation component of the Early Language and Literacy Classroom Observation (ELLCO) (Halle and Vick 2007).

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: CREP trainers conduct group sessions to ensure that observers learn to identify and code variables consistently. They provide observers with manuals, conduct practice exercises, and facilitate an inter-rater reliability/consensus-rating process among observers (Halle and Vick 2007).

Adaptations/Special Instructions for Individuals with Disabilities: Not applicable.

Alternate Forms: Not applicable.

Previous Version: None.

NCEE or REL Study Use:[3] An Impact Evaluation of Early Literacy Programs: The Effects of Opening the World of Learning (OWL) on the Early Literacy Skills of At-Risk Urban Preschool Students (REL-Appalachia)

[1] If researchers and personnel from the Center for Research in Educational Policy conduct observations, then users do not pay training costs (manuals, training session, and trainer's travel). S. J. Hurst provided cost information (personal communication, October 28, 2008).

[2] Reliability ratings for inter-rater reliability are only available for the E-LOT.

[3] See Table F.1 for web address.

References:

Grehan, Anna, and Steven M. Ross. "An Evaluation of the Effects of FOCUS on First Grade Reading Achievement in a Title I Elementary School." Memphis, TN: Center for Research in Educational Policy, University of Memphis, 2004.

Grehan, Anna, Steven M. Ross, and Güler Boyraz. "Reading First: Year 2 Evaluation Results." Paper presented at the annual meeting of the American Educational Research Association, Chicago, 2007.

Grehan, Anna, Lana Smith, Güler Boyraz, Ying Huang, and Debbi Slawson. "Tennessee Reading First Formative Evaluation." Memphis, TN: Center for Research in Educational Policy, University of Memphis, 2006.

Grehan, Anna, and Allan Sterbinsky. "Literacy Observation Tool Reliability Study." Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada, 2005.

Grehan, Anna, Lisa Dyson, and Lana J. Smith. *Adolescent Literacy Observation Tool (A-LOT)*. Memphis, TN: Center for Research in Educational Policy, University of Memphis, 2007.

Grehan, Anna, and Lana J. Smith. *The Early Literacy Observation Tool (E-LOT)*. Memphis, TN: Center for Research in Educational Policy, University of Memphis, 2004.

Halle, Tamara, and Jessica Vick. "Quality in Early Childhood Care and Education Settings: A Compendium of Measures." Submitted to the Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Washington, DC: Child Trends, November 2007.

Huang, Ying, Louis Franceschini, and Steven M. Ross. "Inter-Rater Reliability Analysis of E-LOT." Memphis, TN: Center for Research in Educational Policy, University of Memphis, 2007.

Smith, Lana J., Steven M. Ross, and Anna Grehan. *Literacy Observation Tool (LOT)*. Memphis, TN: Center for Research in Educational Policy, University of Memphis, 2002.

OBSERVATION MEASURE OF LANGUAGE AND LITERACY INSTRUCTION (OMLIT), 2006

Authors: Barbara D. Goodson, Carolyn J. Layzer, and W. Carter Smith		**Type of Assessment:** Classroom observation **Domain:** Classroom quality; instructional practices (reading)
Publisher: Abt Associates 617-492-7100 http://www.abtassociates.com		**Grade/Age Range:** Early childhood classrooms **Administration Interval:** As frequently as desired
Material, Training, and Scoring Costs: Materials: Copies of the training manual and assessment materials are available online from the publisher Training: $1,000 a day plus expenses for an official trainer[1]		**Personnel and Training Requirements** **Credentials Required for Use:** Not specified **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours)
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 4 (>$500) **Time to Administer:** Minimum 2.5 hours **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3[2] (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Observation Measure of Language and Literacy Instruction (OMLIT) is a classroom observation measure used to describe the quality of instructional practices and environment in early childhood classrooms that support language and literacy skill development for students who are dual language learners. The OMLIT was initially developed for the Even Start Classroom Literacy Interventions and Outcomes (CLIO) Study. Observations should last one-half day or longer. The OMLIT consists of six observation tools, with administration recommended in the following order: Classroom Description, Snapshot of Classroom Activities (Snapshot) (completed every 10 minutes), the Classroom Literacy Instruction Profile (CLIP) (completed every 10 minutes, 5 minutes after the Snapshot), Real-Aloud Profile (RAP) (completed any time the adult begins to read aloud to a specified number of students), Quality of Language and Literacy Instruction (QUILL) (based on evidence from all other measures and other events during the day), the Classroom Literacy Opportunities Checklist (CLOC), and the addition of information to the Classroom Description. No scheduled Snapshot or CLIP is completed during a RAP; the observer instead makes a note that it was not completed due to an ongoing RAP. All observations are recorded on paper forms with pencil or pen.

The Classroom Description is a record of classroom context, including number and age of students enrolled and present, staff present, language spoken by students, and languages spoken by staff in instruction. It is broken into six sections. The first four are completed before the observation and include the setting profile, list of staff, description of students (number and age of students, languages spoken at home, and special needs), and classroom themes. The last two sections (languages of instruction and atypical observation event [fire drill, emergency, and so forth]) are completed after the entire observation.

The Snapshot serves as a time sample that provides a picture of classroom activities and the number of students and adults involved in each activity every 15 minutes. It is split into two sections, the first describing the number of students and adults present, and the second describing the activities taking place, including the number of students and adults involved in each activity, languages spoken, and literacy resources used. Any Snapshots scheduled to be administered during a RAP are skipped.

The CLIP is also a time sample that provides a description of literacy activities in the classroom and instructional methods used by the staff. At predetermined intervals (every 15 minutes), the observer determines if a literacy activity is ongoing. If so, the observer records details of the activity for the next 10 minutes along seven characteristics: type of activity, area of literacy, teacher's instructional style, text support, languages spoken, interaction partners,[3] and number of students. If discussion among the adult and students is involved, the observer rates quality on a 5-point scale (from 1 = minimal to 5 = high with item-specific descriptors) for two characteristics: cognitive challenge (including cognitive abstraction and cognitive extension) and depth of discussion. If the teacher is not engaged in any literacy activity, the observer determines if the teacher's aide is engaged. If no instructor is engaged in a literacy activity at the time of the scheduled CLIP, the observer records details of non-literacy activities. Any CLIP scheduled to be administered during a RAP are skipped.

The RAP is an event sample that provides a description of the instructional practices used when staff read aloud to students, with a focus on dialogic reading practices, using a total of 48 items. The observer records adult behavior in four categories: pre-reading set-up behavior, behavior while reading, post-reading behavior, and language used while reading. The environment of the read-aloud is characterized in three categories: role of the adult reading, number of students in the read-aloud, and characteristics of the book read. In addition, observers record the quality of the read-aloud on a 5-point scale (from 1 = minimal to 5 = high with item-specific descriptors) for three categories: the introduction of story-related vocabulary, the extent to which the adult uses open-ended questions to engage the students, and depth of post-reading book-related activities organized by the adult.

The QUILL rates the quality of 10 practices in six areas of literacy instruction and support for language and literacy development in the classroom, specifically including those activities or practices that increase students' oral language skills, vocabulary development, phonological awareness, letter and work knowledge, print awareness, and writing skills. Observers rate the quality of implementation for 10 practices on a 5-point scale (from 1 = minimal to 5 = high with item-specific descriptors). Six of the practices are also rated with a 4-point frequency scale. Four practices are relevant only for activities including English Language Learners and do not include a frequency scale. Four additional items (beyond items for the 10 practices used in the CLIO study) used in the earlier version address opportunities for students with special education needs as well as opportunities to display interests in print, creative arts, and dramatic play.

The CLOC is an inventory of the adequacy of classroom literacy resources available to students. Based on the entire half-day observation, observers rate the classroom from 1 (insufficient or minimal) to 3 (sufficient or high) on 55 classroom literacy resources. The items are grouped into categories for classroom physical layout, print environment, books and reading area, listening area, writing resources, cultural and linguistic diversity in materials, literacy materials and toys within and outside reading and writing areas, instructional technology, richness and integration of a curriculum theme, and literacy resources outside the classroom.

Other Languages: None.

Uses of Information: The OMLIT is used to assess the quality of the environment and instructional practices intended to support the development of language and literacy skills in early childhood classrooms that include students who are English Language Learners.

Methods of Scoring: Counts of materials and practices and ratings of frequency and quality are made according to the rubrics that accompany the observation tools. Given the different types of data collected, the CLIO study transformed the rating or count on each individual behavior into a standard score with a mean of 0 and standard deviation of 1 and created five theoretical measures with items drawn from across the components within the OMLIT: support of oral language (14 items), support for phonological awareness (4 items), support for print knowledge (16 items), support for print motivation (5 items), and adequacy of literacy resources (7 items).

Interpretability: A higher raw score could indicate greater quantity or quality depending on the scale used. To increase interpretability, the CLIO study aggregated classroom raw scores to the project level and rescaled raw scores into *T*-scores (M = 50, SD = 10) using the 2004 control

group mean as the referent group. Thus, those above the mean demonstrated the practices more than the control group (i.e., those without any intervention), and those below the mean implemented them less frequently than the mean of the control group.

Reliability:
(1) Internal consistency reliability: An internal consistency approach for estimating the reliability of scores was conducted for both the CLIO sample (N = 199 classrooms) and a sample in the Miami Child Care Study (N = 162 classrooms). Reliability estimates ranged from 0.72 to 0.84 for each theoretical scale, except for phonological awareness, which ranged from 0.58 to 0.61.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: Not applicable.
(4) Inter-rater reliability: During training, observers practiced coding on criterion-referenced paper-and-pencil tests and on videotaped classroom scenarios. Only those with 75 percent agreement on both practices went on to perform field observations. Inter-rater reliability estimates based on field observations ranged from 67 to 98 percent agreement on individual behaviors. These statistics are based on exact agreement between paired observers in 90 observations over three waves of data collection. In addition, for the five measures derived from the OMLIT battery for the CLIO study, the authors calculated percentage exact agreement on the measure scores based on 33 paired observations. For the five measures, correlations ranged from 0.80 to 0.89.

Validity Evidence:
Each measure's development was based on research into instructional practices that have been shown to predict students' reading skills and other academic outcomes and on research into the environmental influences, such as behavior and resources that affect language development. Earlier versions of the instruments were piloted in fall 2003 in six classrooms in child care facilities. Based on the results, the measures were revised as follows: deletion of poorly functioning items; combining related items that led to unreliable information individually but that together resulted in reliable information; the addition of new items to address gaps in the measure; and revision of training materials to provide clearer definitions among codes. The CLIO study examined the intercorrelations among the five theoretical measures it developed from ratings on the OMLIT tools (see Method of Scoring). Adequacy of literacy resources correlated with print knowledge at 0.25. Support for phonological awareness was not related to support for oral language. The four support constructs' intercorrelations ranged from 0.15 to 0.39.

Construct/Concurrent validity: Without adjusting for multiple comparisons, the CLIO study detected significant differences between intervention and control groups on four of the five measures derived from the OMLIT (the exception was Support for Oral Language Development). After adjusting for multiple comparisons, significant differences were found only for support for print knowledge and literacy resources in the classroom. However, in examining relationships between instructional measures and student outcomes after controlling for other parent and teacher behaviors, only support for print knowledge was positively related to any of the student language and literacy outcomes; both oral language and support for print knowledge were negatively related to outcomes on the Spanish Individual Growth and Development Indicator (IGDI).

Predictive validity: No information available.

Bias analysis: No information available.

Training Support: The developer offers training in the measure at a cost of $1,000 a day plus expenses per trainer. The Snapshot, RAP, CLIP, and QUILL require eight hours of training each, and the CLOC and Classroom Description require less than a half-day of additional training. Two types of training are given: classroom training, with inter-rater reliability tests; and practice observation in a preschool classroom, ideally also including inter-rater reliability tests between trainee and trainer.

Adaptations/Special Instructions for Individuals with Disabilities: Not applicable.

Alternate Forms: None.

Previous Version: The OMLIT was originally developed in 2003 and revised twice in 2004. Revisions included the deletion of poorly functioning items; combining related items that led to unreliable information individually but that together resulted in reliable information; the addition of new items to address gaps in the measure; and revisions to training materials to provide clearer definitions among codes. Substantial changes were also made to the QUILL and CLOC.

NCEE or REL Study Use:[4] Even Start Classroom Literacy Interventions and Outcomes (CLIO)

[1] Halle and Vick 2007.

[2] Certain items of the OMLIT battery of measures encompassed some ratings below the 0.70 level (see Reliability).

[3] Observers indicate if the child is talking with the teacher, with peers, or with the group and note if the teacher's language is directed toward a group, one child, or children in turn.

[4] See Table F.1 for web address.

References:

Goodson, Barbara D., and Carolyn J. Layzer. "Assessing Support for Language and Literacy in Early Childhood Classrooms: The Observation Measures for Language and Literacy (OMLIT: Goodson, Layzer, Smith, and Rimdzius, 2004)." Paper presented at the annual conference of the American Educational Research Association, Montréal, 2005.

Goodson, Barbara D., Carolyn J. Layzer, W. Carter Smith, and Tracy Rimdzius. *Observation Booklet Spring 2006, Observation Measures of Language and Literacy Instruction (OMLIT)*. Cambridge, MA: Abt Associates, 2004, 2006.

Goodson, Barbara D., Carolyn J. Layzer, W. Carter Smith, and Tracy Rimdzius. *OMLIT Training Manual, Chapters 1 and 2*. Cambridge, MA: Abt Associates, 2006.

Halle, Tamara, and Jessica Vick. "Quality in Early Childhood Care and Education Settings: A Compendium of Measures." Submitted to the Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Washington, DC: Child Trends, 2007.

Judkins, David, Robert St. Pierre, Babette Gutmann, Barbara D. Goodson, Adrienne von Glatz, Jennifer Hamilton, Ann Webber, Patricia Troppe, and Tracy Rimdzius. "A Study of Classroom Literacy Interventions and Outcomes in Even Start." Report prepared for the Institute of Education Sciences. (NCEE 2008-4028). Washington, DC: National Center for Education Evaluation, 2008.

Layzer, Jean I., Carolyn J. Layzer, Barbara D. Goodson, and Cristofer Price. "Evaluation of Child Student Care Subsidy Strategies: Findings from Project Upgrade in Miami-Dade County." Submitted to the U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation and The Child Care Bureau. Cambridge, MA: Abt Associates, 2007.

REFORMED TEACHING OBSERVATION PROTOCOL (RTOP), 2000

The Reformed Teaching Observation Protocol (RTOP) is also a teacher knowledge measure and thus is found under Appendix C, Teacher Knowledge Measures. Please refer to Appendix C for this profile.

SCHOOL OBSERVATION MEASURE (SOM), 1999

Authors: Steven M. Ross, Lana J. Smith, and Marty Alberg		**Type of Assessment:** Classroom observation **Domain:** Classroom quality (environment), instructional practices (comprehensive), school climate, school engagement
Publisher: Center for Research in Educational Policy (CREP) University of Memphis 866-670-6417 http://www.crep.memphis.edu/		**Grade/Age Range:** Kindergarten through grade 12 **Administration Interval:** As frequently as desired
Material, Training, and Scoring Costs: Copies of the SOM are available through CREP.[1] Use of the measure requires signing a contract and completing CREP training. Training (excluding travel, supplies, venue): $2,000 per day for up to 40 attendees		**Personnel and Training Requirements** **Credentials Required for Use:** Level 1 (training or supervised experience with measurement) **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (> 2 hours)
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (under $100) **Time to Administer:** 3 hours **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3[2] (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: Ross et al. (1999) developed the School Observation Measure (SOM) as a tool for assessing the instructional practices used in an elementary, middle, or high school. Observers summarize the instructional practices they observe during 10 15-minute observations conducted in several classrooms in one school in one day. The measure is designed to provide a school-wide perspective on the prevalence of 24 teaching strategies or events that may be used separately or concurrently during classroom instruction. The strategies, each of which is assessed via a single item on the measure, are classified according to six categories: (1) Instructional Orientation (four items); (2) Classroom Organization (three items); (3) Instructional Strategies (six items); (4) Student Activities (seven items); (5) Technology Use (two items), and (6) Assessment (two items). In addition to the 24 classroom strategies or events, the SOM includes two summary items pertaining to the degree of academic focus and student engagement observed during a lesson. During the classroom observations, observers use a Notes Form to indicate whether they observed each of the 24 strategies and to write comments about what they observed. After the school visit, the observer completes an SOM Data Summary Form to rate the frequency across all classrooms in a school for the 24 strategies (5-point scale; not observed to extensively) and the two summary items on academic focus and student engagement (3-point scale; low to high).

The developers recommend a minimum number of classrooms and observations to ensure adequate assessment of practices and setting at the school level. They recommend that observers conduct 15-minute observations in 10 to 12 classrooms during a three-hour visit to a school. Classroom selection should be random while allowing for observation of classrooms in different grades and minimizing repeat observations of the same classrooms over time. Observers should observe regular classes (i.e., not involving groups of selected students, such as special education or English as a Second Language classes) focused on core academic subjects. To provide adequate data to characterize a school's instructional practices, the developers recommend that observers conduct observations (as described above) at least 6, but preferably 8 to 10, times per school year. Observations should be spread across times of day, days of the week, and weeks of the school year. Individual observers generally conduct observations, except in cases of paired observations for purposes of assessing inter-rater reliability. To avoid observer bias effects, at least two observers should conduct observations in a school.

Other Languages: None.

Uses of Information: Researchers and program evaluators may use the SOM to assess the frequency and prevalence of instructional practices on a school-wide basis. The developers note that they developed the SOM as a formative evaluation tool to provide information on the degree to which schools involved in whole-school reform implemented targeted instructional goals.

Methods of Scoring: To complete a SOM Data Summary Form, the observer reviews his or her class-specific notes and then rates the frequency with which he or she observed, across all observed classrooms in the school, each of the 24 instructional strategies or events that

are assessed as single items on the measure. Each item consists of an item stem and a five-point rubric (0 = not observed, 1 = rarely, 2 = occasionally, 3 = frequently, and 4 = extensively). The two summary items (academic focus, student engagement) have a three-point rating scale (1 = low, 2 = moderate, and 3 = high). According to the developers, the selection of a rating requires observers to make a holistic judgment about a strategy or event based on (1) the number of classrooms in which they observed it and (2) their perception of its prevalence or the amount of emphasis it received. In descriptive and inferential data analyses, SOM results may be interpreted as frequencies and/or converted to ordinal scores. A summary report, aggregating frequency data for each item across all classrooms in a school over several observation visits, provides information about school-wide instructional practices over time.

Interpretability: Ross et al. (2004) stress that SOM results indicate the frequency and prevalence of selected teaching strategies across classrooms in a school but do not indicate the quality of such strategy implementation. Users of the SOM may review SOM findings for individual classrooms, but the authors state that the measure was designed to yield school-level data for use in informing school-wide professional development and curricular planning.

Reliability:
(1) Internal consistency reliability: No information available.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: Not applicable.
(4) Inter-rater reliability: In a study of inter-rater reliability conducted by Ross et al. (1999), 10 pairs of trained observers, each consisting of one expert observer, conducted joint observations of approximately 10 classrooms in Memphis. Each observer independently completed the Notes Forms and Data Summary Forms. The authors conducted difference scores analyses to establish the percentage of times the two observers agreed in their ratings or disagreed by one, two, or more points on the rubric. Observers achieved perfect agreement on 67.7 percent of the ratings and agreed within one category 93.8 percent of the time. They agreed within two categories 100 percent of the time. Bivariate correlations for each pair of observers across all items ranged from 0.68 to 0.97, with a median of 0.76. The authors reported item analyses demonstrating high inter-rater agreement on all but five items: (1) cooperative learning, (2) ability grouping, (3) higher-level instructional feedback, (4) sustained writing, and (5) teacher acting as coach or facilitator.

Validity Evidence:
The developers of the SOM based the item content on national teaching standards and teaching methods associated with contemporary educational reforms and empirically linked to improved academic achievement. Twelve of the items had been used in a previous measure, the Classroom Observation Measure (COM; Ross et al. 1994). A panel of school professionals and educational researchers reviewed and helped refine the items and observational procedures.

Construct/Concurrent validity: The SOM has been used in formative evaluations of the effectiveness of school reform designs. Schools' ratings on the SOM have been consistent

with stated instructional emphases. In an evaluation of the Co-Nect school reform design in five Memphis elementary schools, Ross et al. (2000) performed SOM observations in matched treatment and control schools. The reform design called for the treatment schools to emphasize technology use, project-based learning, and student engagement in active learning activities. The authors reported that observers using the SOM observed significantly more project-based learning, use of computers as a learning tool, use of computers for instruction, sustained writing, and independent inquiry in treatment schools as compared to control schools. Ross et al. (2000) also used the SOM in a study of the effects of providing grade 5 and 6 students with laptop computers at home and school. The authors collected SOM data in targeted groups of 32 treatment (laptop) and 18 control classrooms during 60-minute observations (rather than the typical whole-school survey method involving 15-minute observations in randomly sampled classrooms within a school). Observers noted significantly more project-based learning, independent inquiry/research, and computer use as a learning tool and for delivering instruction in treatment versus control classrooms.

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: The developers require observers to complete formal training before using the SOM. At a one-day training session conducted at the Center for Research in Educational Policy at the University of Memphis, trainees receive a detailed training manual. After a discussion of the observation and rating procedures, trainees conduct practice observations by using videotaped simulations. The training also includes a 1.5-hour practice observation at a school. Groups of two to four observers conduct five 15-minute observations, followed by a whole-group review session with a trainer. Observers are required to conduct their first post-training SOM with a partner, with both observers independently completing the rating form, comparing ratings, and completing a "consensual" form from which inter-rater reliability estimates may be determined. Observers may consult with CREP trainers by telephone or email for clarification or feedback.

Adaptations/Special Instructions for Individuals with Disabilities: Not applicable.

Alternate Forms: Not applicable.

Previous Version: None.

NCEE or REL Study Use:[3] The Effects of Hybrid Algebra I on Teacher Practices, Classroom Quality, and Adolescent Learning

[1] D. Strahl, personal communication, February 4, 2009.

[2] Inter-rater reliability coefficients met a minimum level of acceptability ($\geq$0.70). Percentages of agreement between raters met a minimum level of acceptability (85 percent) within one point on the rubric; however, exact agreement percentages were below the minimum level.

[3] See Table F.1 for web address).

References:

Ross, Steven M., and Deborah L. Lowther. "Impacts of the Co-Nect School Reform Design on Classroom Instruction, School Climate, and Student Achievement in Inner-City Schools." *Journal of Education for Students Placed at Risk*, vol. 8, no. 2, 2003, pp. 215-246.

Ross, Steven M., Deborah L. Lowther, Robert T. Plants, and Gary R. Morrison. *Final Evaluation of the Anytime, Anywhere Learning Laptop Program*. Memphis, TN: Center for Research in Educational Policy, University of Memphis, 2000.

Ross, Steven M., Lana J. Smith, and Marty Alberg. *The School Observation Measure*. Memphis, TN: Center for Research in Educational Policy, University of Memphis, 1999.

Ross, Steven M., Lana J. Smith, Marty Alberg, and Deborah Lowther. "Using Classroom Observation as a Research and Formative Evaluation Tool in Educational Reform: The School Observation Measure." In *Observational Research in U.S. Classrooms: New Approaches for Understanding Cultural and Linguistic Diversity*, edited by Hersh Waxman, Roland G. Tharp, and R. Soleste Hilberg. New York: Cambridge University Press, 2004.

Ross, Steven M., Lana J. Smith, Linda Lohr, and Mary McNelis. "Math and Reading Instruction in Tracked First-Grade Classes." *The Elementary School Journal*, vol. 95, no. 1, 1994, pp. 105-109.

SHELTERED INSTRUCTION OBSERVATION PROTOCOL (SIOP), 2008

Authors: Jana Echevarria, MaryEllen Vogt, and Deborah J. Short		**Type of Assessment:** Classroom observation **Domain:** Instructional practices (Sheltered Instruction for English Language Learning), classroom quality
Publisher: Pearson Assessments 800-627-7271 http://www.siopinstitute.net		**Grade/Age Range:** Kindergarten through Grade 12 **Administration Interval:** As frequently as desired
Material, Training, and Scoring Costs: The SIOP form is published in *Making Content Comprehensible for English Learners: The SIOP Model, 3rd Edition* (Echevarria et al. 2008); reproducible with purchase of the book ($44.09)		**Personnel and Training Requirements** **Credentials Required for Use:** No special requirements required **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours)
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (under $100) **Time to Administer:** See description **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70)[1] **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available[2] **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Sheltered Instruction Observation Protocol (SIOP) is an observational rating scale for scoring teachers' implementation of the Sheltered Instruction (SI) model of teaching language and subject matter content to English Language Learner (ELL) students. SI combines language and content objectives into subject area curricula, with teachers presenting grade-level subject matter through modified instruction in English to enhance comprehension. The protocol may be used to observe sheltered instruction with ELLs ranging from beginning to advanced in proficiency. While the authors do not specify grades for which the SIOP is appropriate the SIOP websites (http://www.siopinstitute.net; http://www.cal.org/siop/) note use by elementary and secondary teachers.

The protocol consists of 30 items that operationalize key features of sheltered instruction. The items are grouped into eight main subscales (referred to by the authors as components): Lesson Preparation (using clearly defined objectives, meaningful materials, and lesson planning activities); Building Background (activating previous knowledge and building academic vocabulary); Comprehensible Input (clear speech and explanations with multimodal approaches); Strategies (teaching specific learning strategies, scaffolding, and building higher-order thinking skills); Interaction (promoting extended interactions about concepts and giving students adequate time to respond to questions); Practice/Application (providing activities to practice and extend learning); Lesson Delivery (student engagement, pacing, and delivery that supports objectives); and Review/Assessment (reviewing key concepts, assessing mastery, and providing feedback to students).

The SIOP may be completed by observers during the observed lesson and/or after the lesson with observation notes or while viewing a videotape of the lesson. Observers may use the full form or abbreviated form, both of which have the same 30 items noted above. The forms differ in format and number of pages (the full and abbreviated forms are six and two pages in length, respectively). Both forms require observers to rate each item on a 5-point scale to indicate the degree to which a feature is evident. On the full form, item content is embedded into the response scale, with item-specific descriptors at the endpoints and middle of the 5-point scale. On the abbreviated form, each item consists of an item stem with a brief response format that is uniform across items (from not evident to highly evident). On both forms, some items may be marked NA (Not Applicable) if a feature or behavior is not relevant to a lesson. Observers may also record qualitative comments about behaviors or practices that were or were not demonstrated. Both forms yield an overall score that is the ratio between the sum of scores across all items and the total possible score, which is converted into a percentage.

Other Languages: None.

Uses of Information: The SIOP instrument may be used to measure fidelity to the SIOP model. It may also be used to provide feedback to teachers on their implementation of the SIOP model and as a planning and self-assessment tool. Users are cautioned not to rely on ratings of single observations for evaluating teachers' implementation of the SIOP model. Periodic ratings across time not only increase the reliability of the information obtained but can document growth in teachers' performance.

Methods of Scoring: After recording a rating for each item (from 0 to 4 if the feature or behavior is relevant to the lesson or NA if it is not), the observer sums the scores across all 30 items. The sum is then divided by the total possible score (usually 120; but for each NA rating, 4 points are subtracted from 120), and the resulting proportion is multiplied by 100 to create a percentage score. Scoring instructions do not include procedures for calculating subscale scores (Echevarria et al. 2008).

Interpretability: SIOP scores and corresponding percentages are meant to serve as an indicator of the level of implementation of the SIOP model. There are neither norms nor guidelines for interpreting the scores. The scores provide a basis for collaborative discussion among teachers and colleagues, supervisors, or trainers. The authors state that plotting scores for each item on a line graph can highlight areas of strength as well as potential focus areas for further practice or training. For practice and research purposes, longitudinal observations and scores may be used to measure improvements or fidelity in SI model implementation over time. The authors caution against drawing conclusions from single observations in that many variables can influence implementation of a single lesson. They recommend rating several lessons over time for a more accurate assessment of teachers' SI implementation.

Reliability:
(1) Internal consistency reliability: Internal consistency information is not available for the current version of the SIOP with a 5-point scale and 30 items. Echevarria et al. (2007) summarized findings from a field test for a preliminary version that consisted of 31 items, each with a 7-point response format. They reported Cronbach's alphas ranging from 0.87 to 0.96 for scores from eight theoretically derived subscales (Preparation, Building Background, Comprehensible Input, Strategies, Interaction, Practice/Application, Lesson Delivery, and Review/Evaluation). For a revised version (30 items each with a 7-point scale), Guarino et al. (2001) reported alpha coefficients for scores on three subscales derived from principal components analysis: Preparation (6 items), Instruction (20 items), and Review/Evaluation (4 items). The alpha coefficients for scores on these subscales ranged from 0.92 to 0.98. Even though the SIOP is designed to yield a total score, neither investigation reported an internal consistency reliability estimate for a total score.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: Not applicable.
(4) Inter-rater reliability: Inter-rater reliability information is not available for the current version of the SIOP. Echevarria et al. (2008) reported that, in a field test of a preliminary version of the SIOP that had a 7-point response format, university-based teacher supervisors who had not been trained in the SIOP model were able to distinguish between middle school teachers demonstrating high and low implementation of the model, with an inter-rater correlation coefficient of 0.99.

Validity Evidence:
The authors and collaborating teachers developed the SIOP items to reflect the key components and characteristics of the SIOP model. The model and corresponding observation protocol were based on the best-practices research literature in teaching English as a Second Language, bilingual education, reading, language and literacy acquisition, discourse studies, special education, and classroom management as well as on the developers' own knowledge and experience. Early versions of the

SIOP were pilot-tested with samples in four large urban school districts (two East Coast and two West Coast) for classrooms with beginning to advanced ELL students.

Construct/Concurrent validity: Echevarria et al. (2007) cited results of a principal component analysis with varimax rotation on a preliminary version of the SIOP that had 31 items and a 7-point response format. The analysis yielded three components--Preparation, Instruction, and Review/Evaluation—that explained 98.4 percent of the variance.

Echevarria et al. (2006) found that SIOP ratings related in expected ways to student outcomes. Using a version of the SIOP protocol that had 30-items and a 5-point response scale, they found that middle school ELL students of teachers who were trained in the SIOP model and received high scores on the SIOP measure exhibited higher posttest writing scores, and greater gains in writing scores, than students whose teachers did not receive training in the SIOP model.

As for discriminant analysis, researchers performed discriminant functional analysis (DFA) on a preliminary version of the SIOP (with 31 items and a 7-point response scale) to discriminate between teachers who did and did not perform SI (as cited in Echevarria et al. 2007). Using the eight subscales as predictors, they calculated one significant discriminant function. Univariate tests indicated that the subscales that best discriminated between SI and non–SI teachers were Preparation, Lesson Delivery, Comprehensible Input, Building Background, Strategies, Practice/Application, and Review/Evaluation. Given that the Interaction subscale did not discriminate between the two groups, its respective items were modified (and one was dropped) in a subsequent SIOP revision. In a discriminant analysis of the revised version (30-items, 7-point response scale) with three subscales (Preparation, Instruction, and Review/Evaluation) as predictors of instruction types (SI and non–SI), (Guarino et al. (2001) calculated one significant discriminant function. Univariate tests showed that all three predictors discriminated between the two groups.

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: Training and materials supporting the use of the SIOP model of teaching ELLs is available through two sources: (1) The SIOP Institute (Pearson Education, Inc., http://www.siopinstitute.net) and the Center for Applied Linguistics (CAL; http://www.cal.org/siop/). The training and instructional materials focus on the SIOP teaching methods in general, of which use of the SIOP measure is a subtopic. The SIOP Institute offers training seminars for individuals and on-site group training for school districts as well as several training manuals that include DVDs. CAL also offers training manuals and DVDs and professional development services for groups (including workshops, coaching, site visits, and technical assistance to schools and districts).

Adaptations/Special Instructions for Individuals with Disabilities: Not applicable.

Alternate Forms: Not applicable.

Previous Version: The SIOP form published in *Making Content Comprehensible for English Learners: The SIOP Model, 3rd Edition* (Echevarria et al. 2008) incorporates slight changes in graphics and item wording.

NCEE or REL Study Use:[3] Evaluation of Principles-Based Professional Development to Improve Reading Comprehension for English Language Learners

[1] Reliability estimates are not available for the current version of the SIOP protocol. Studies conducted with preliminary versions reported alphas greater than 0.70 (see Reliability).

[2] Most of the validity studies used earlier versions of the SIOP protocol that had either 31 or 30 items and 7-point response scales (see Construct/Concurrent validity).

[3] See Table F.1 for web address.

References:

Echevarria, Jana, and Deborah Short. "Using Multiple Perspectives in Observations of Diverse Classrooms: The Sheltered Instruction Observation Protocol (SIOP)." In *Observational Research in U.S. Classrooms*, edited by Hersh Waxman, Roland Tharp, and R. Soleste Hilberg. Cambridge, UK: Cambridge University Press, 2004.

Echevarria, Jana, Deborah Short, and K. Powers. "School Reform and Standards-Based Education: An Instructional Model for English Language Learners." *Journal of Educational Research*, vol. 99, no. 4, 2006, pp. 195-211.

Echevarria, Jana, Deborah Short, and MaryEllen Vogt. *Implementing the SIOP Model through Effective Professional Development and Coaching*. Boston: Pearson Allyn & Bacon, 2007.

Echevarria, Jana, MaryEllen Vogt, and Deborah Short. *Making Content Comprehensible for English Learners: The SIOP® Model (3rd Ed.)*. Boston: Pearson Allyn and Bacon, 2008.

Guarino, A.J., Jana Echevarria, Deborah Short, J.E. Schick, S. Forbes, and R. Rueda. "The Sheltered Instruction Observation: Reliability and Validity Assessment." *Journal of Research in Education*, vol. 11, no. 1, 2001, pp. 138-140. Pre-publication copy obtained from author.

TEACHER BEHAVIOR RATING SCALE (TBRS), 2004

Authors: Susan H. Landry, April Crawford, Susan Gunnewig, and Paul R. Swank	**Type of Assessment:** Classroom observation **Domains:** Classroom quality (teacher-student interactions and classroom environment) and instructional practices (comprehensive)
Publisher: Unpublished measure developed by the Center for Improving the Readiness of Children for Learning and Education (CIRCLE) 713-500-3714 http://www.childrenslearninginstitute.org/our-programs/program-overview/CIRCLE/	**Grade/Age Range:** Preschool **Administration Interval:** As frequently as desired
Material, Training, and Scoring Costs: Not specified	**Personnel and Training Requirements** **Credentials Required for Use:** Not specified **Personnel for Administration:** Highly trained individual **Training for Administration:** Extensive (>2 hours) Observers must complete field certification before conducting independent observations.
Languages: English	**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample	**Summary** **Initial Material Cost:** To be determined upon negotiation with publisher **Time to Administer:** 2 to 3 hours **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3[1] (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available[1] **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Teacher Behavior Rating Scale (TBRS) is a classroom observation measure designed to assess the overall classroom environment and specific language and literacy practices of preschool teachers. The scale was adapted most recently for the U.S. Department of Education's National Evaluation of Early Reading First (ERF) study and comprises approximately 60 items across the following 17 subscales: Classroom Community, Teacher Sensitivity, Lesson Planning, Child Portfolios, Dynamic Assessments, Quality and Organization of Activity Centers, Book-Reading Practices, Number of Phonological Awareness Activities Observed, Quality of Phonological Awareness Activities, Oral Language Use by Lead Teacher, Print and Letter Knowledge Learning Opportunities, Classroom Print Environment, Written Expression Learning Opportunities, Opportunities and Materials for Writing, Math Concepts, Quality of Team Teaching, and Oral Language Use by Assistant Teacher. The number of items per subscale ranges from 1 (Written Expression Learning Opportunities and the phonological awareness subscales) to 8 (Book-Reading Practices). The ratings are based on an observation at least two hours in length that ideally includes only cognitive-based activities. In addition to the classroom observation component, the observer reviews the teacher's lesson plans and student portfolios (Assel et al. 2007). For the majority of items, the observer records both a quantity and quality rating. The quantity score measures either how frequently a behavior occurs or the number of examples observed (none to often). The quality score considers the nature of the behavior or environment, ranging from low to high quality (Jackson et al. 2007).

Other Languages: None.

Uses of Information: The measure is designed to serve (1) as an outcome measure to assess the quality of the classroom environment and (2) as a tool to help guide teachers' practice.

Methods of Scoring: For the majority of the subscales, items are rated in terms of both the quantity and quality of the behavior or environment. The ERF version of the measure bases both quantity and quality ratings on a four-point scale; quantity options are "none," "rarely," "sometimes," and "often" while quality is scored as "low," "medium-low," "medium-high," or "high." The developer provides detailed scoring instructions for each item. Given the high correlation between the quantity and quality scores, the quantity and quality scores for each item are averaged to create a single item score (Jackson et al. 2007). The observer may compute scores for the individual subscales and a total score. The observer creates subscale scores by averaging the item scores for all items within that subscale. The total TBRS score is the average of the 17 subscale scores.

Interpretability: Subscale scores describe specific classroom instructional practices or environmental characteristics while the total score provides an overall picture of the quality of the classroom environment, interactions, and instruction. Given the separate collection of quantity and quality information, separate scores may be created or, as with the ERF, may be averaged. In the latter case, a low score indicates a classroom with infrequent, low-quality interactions or environmental characteristics while a high score points to desired, high-quality instructional practices or environmental characteristics that occur frequently. The interpretation of middle-range scores (mid-level-quality behaviors that occur sometimes) becomes more

difficult if the averaged quantity and quality scores demonstrate correlations below 0.90 (Jackson et al. 2007).

Reliability:[1]
(1) Internal consistency reliability: The U.S. Department of Education's Preschool Curriculum Evaluation Research (PCER) project used the TBRS. The correlations between the quantity and quality item ratings across subscales ranged from 0.72 to 0.97, and Cronbach's alphas for scores based on the combined quality and quantity ratings ranged from 0.82 to 0.95 (Jackson et al. 2007). In the ERF study, the correlation between the quantity and quality item ratings ranged from 0.66 to 0.98 (Jackson et al. 2007).

The internal consistency of scores for the subscales from the original version of the TBRS that provided the basis for the ERF ranged from 0.86 to 0.94, with the exception of the Child Portfolios subscale scores (0.66) and the Dynamic Assessments subscale scores (0.72) (Jackson et al. 2007). The internal consistency of the TBRS subscale scores from the PCER project (Jackson et al. 2007) ranged from 0.63 to 0.97 across evaluations. In the ERF study, the internal consistency for scores across subscales ranged from 0.80 (Classroom Print Environment) to 0.94 (Quality of Team Teaching and Oral Language Use by Assistant Teacher), except for the Child Portfolios subscale scores (0.66) and Dynamic Assessments subscale scores (0.72).
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: Not applicable.
(4) Inter-rater reliability: A 1999 Head Start study piloted the TBRS. Mentors providing professional development were trained on the TBRS and conducted at least five observations per teacher over the course of the year. Reliability was assessed by a mentor coder and an expert TBRS coder, but the exact number of classrooms or observations was not specified. Overall, the level of inter-rater reliability was estimated at 0.78 (Assel et al. 2007). The inter-rater reliability in the PCER study was estimated at 0.73 (Preschool Curriculum Evaluation Research Consortium 2008). In the ERF study, two observers independently assessed 13 teachers, with inter-rater reliability estimates ranging from 0.75 for the phonological awareness subscale to 1.0 for the Portfolios subscale. The estimated correlation between the two observers' ratings for the total TBRS score was 0.93 (Jackson et al. 2007).

Validity Evidence:[1]
The TBRS's original function was to assess the implementation fidelity of an intervention designed to enhance the early literacy and language instruction by Head Start teachers (Landry et al. 2006). According to the researchers, the then-available classroom observation measures assessed only the quality of caregiving behaviors; thus, the researchers developed the TBRS to assess teachers' use of research-based language and literacy instructional practices in addition to the overall classroom environment. The TBRS was also used with additional prekindergarten programs and revised to reflect more accurately what can be measured in a two- to three-hour observation period (Assel et al. 2007).

Construct/Concurrent validity: In the 1999 Head Start study described above, sensitivity-to-change analyses were conducted, and the authors report that the TBRS was able to capture teachers' growth resulting from professional development services (Assel et al. 2007).

In the same study, the researchers examined the intervention's fidelity of model implementation in that change in teacher instructional behavior occurred as documented by change in TBRS ratings (though no comparison to control teachers was made) as well as by change in student outcomes. Agreement between TBRS ratings and student gains in general was reported as 0.80 (Landry et al. 2006). Student outcome measures included the Peabody Picture Vocabulary Test, Expressive Vocabulary Test, Preschool Language Scale-Auditory Comprehension and Expressive Communication subscales, and the Developing Skills Checklist-Letter Recognition and Auditory subscales. The PCER study also found that scores on the TBRS were significantly positively associated with student gain scores, controlling for age and time between assessments. The Early Childhood Environment Rating Scale–Revised Edition (ECERS-R), a measure of the quality of the classroom environment, was also administered as part of PCER and its association with student outcomes examined. While the TBRS was positively associated with gains, the ECERS-R scores were significantly negatively associated with students' outcomes on language and literacy measures (Assel et al. 2007).

The developers examined data from the PCER study to assess the correlation between TBRS language and literacy subscales and student outcomes assessed with the following measures: Preschool Language Scale-IV (PLS-IV) Auditory Comprehension subscale; the Expressive Vocabulary Test (EVT); the Woodcock-Johnson III Test of Academic Achievement Letter-Word Identification subscale (WJ-III Letter-Word ID); the Woodcock Johnson-III Sound Awareness (rhyming) subscale (WJ-III Sound Awareness); and the Developing Skills Checklist (DSC) Auditory subscale. The correlations for the TBRS Oral Language Use subscale ranged from 0.47 for the DSC to 0.63 for the EVT. For the Phonological Awareness subscale, the correlations ranged from 0.25 for the WJ-III Letter-Word ID to 0.39 for the WJ-III Sound Awareness. The correlations for the Print and Letter Knowledge subscale ranged from 0.37 for the WJ-III Letter-Word ID to 0.55 for WJ-III Sound Awareness. For the General Teaching Behaviors subscale, the correlations ranged from 0.35 for the WJ-III Sound Awareness to 0.57 for the EVT (Assel et al. 2007).

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: Training on TBRS typically involves an overview of the measure, including a review of the constructs underlying the measure, specific subscale coding practice with videotapes, and field certification in which the assessor must achieve a specified percentage agreement calculation with an expert coder (Assel et al. 2007). For the PCER study, a train-the-trainer model was used in which the developers trained representatives from the research organizations who then trained their classroom observers. Training on the TBRS was conducted over four days and included a refresher of two observation measures in the first year of the study (Preschool Curriculum Evaluation Research Consortium 2008). Classroom observers for the ERF study underwent six days of training (covering more than just the TBRS) from the developers and other research organizations involved in the study. The training included the following sessions: study overview, child growth and development, ECERS-R, TBRS, conduct of classroom observations, quality assurance procedures, administrative logistics, and certification. To attain certification, the observers had to achieve an inter-rater reliability of 0.90 (Jackson et al. 2007).

Adaptations/Special Instructions for Individuals with Disabilities: Not applicable.

Alternate Forms: Not applicable.

Previous Version: The TBRS scale was initially developed to assess the fidelity of implementation of an intervention designed by CIRCLE. The scale has been updated and modified over the past few years through its use in various studies. It was first pilot tested in a 1999 Texas study assessing the role of professional development in promoting Head Start students' school readiness (Landry et al. 2006). The scale has since been used in two major studies conducted by the U.S. Department of Education: the PCER study in 2004 and the ERF study in 2005.

The major modification for the PCER study was the development of both a quantity and quality rating for the majority of the items (Jackson et al. 2007). The PCER study included only a subset of the TBRS subscales: written expression, print and letter knowledge, phonological awareness, book reading, oral language, and mathematics concepts. For the ERF study, the quantity and quality ratings were expanded from a three- to a four-point Likert scale, and the quantity and quality scores for each item were averaged to create a single item score. Four of the original TBRS subscales (team teaching, phonological awareness activity, print and letter knowledge, and written expression) were also slightly modified to include an assessment of the assistant teacher's use of language, to simplify scoring, and to create distinct scores for teachers' behavior and the classroom environment's support of the domain.

NCEE or REL Study Use:[2] National Evaluation of Early Reading First

[1] The reliability and validity of the TBRS scores were assessed in three studies, each of which used a different version of the measure. Please see Previous Version for information on how the TBRS was adapted for the studies.

[2] See Table F.1 for web address.

References:

Assel, Mike A., Susan H. Landry, and Paul R. Swank. "Are Early Childhood Classrooms Preparing Children to Be School Ready? The CIRCLE Teacher Behavior Rating Scale." In Achieving Excellence in Preschool Literacy Instruction, edited by Laura M. Justice and Carol Vukelich. New York: Guilford Publications, 2007.

Jackson, Russell, Ann McCoy, Carol Pistorino, Anna Wilkinson, John Burghardt, Melissa Clark, Christine Ross, Peter Schochet, and Paul Swank. "National Evaluation of Early Reading First: Final Report." Submitted to U.S. Department of Education, Institute of Education Sciences. Washington, DC: U.S. Government Printing Office, May 2007.

Landry, Susan H., April Crawford, Susan Gunnewig, and Paul R. Swank. *Teacher Behavior Rating Scale*. Unpublished research instrument. Houston, TX: Center for Improving the Readiness of Children for Learning and Education, University of Texas Health Science Center at Houston, 2004.

Landry, Susan H., Paul R. Swank, Karen E. Smith, Michael A. Assel, and Susan B. Gunnewig. "Enhancing Early Literacy Skills for Preschool Children: Bringing a Professional Development Model to Scale." Journal of Learning Disabilities, vol. 39, no. 4, 2006, pp. 306-324.

Preschool Curriculum Evaluation Research Consortium. "Effects of Preschool Curriculum Programs on School Readiness. Report from the Preschool Curriculum Evaluation Research Initiative." (NCER 2008–2009). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education, 2008.

TEACHER INTERACTION AND LANGUAGE RATING SCALE, 2000

Authors: Luigi Girolametto, Elaine Weitzman, and Janice Greenberg		**Type of Assessment:** Classroom observation **Domain:** Instructional practices (language arts/language proficiency)
Publisher: The Hanen Centre 416-921-1073 http://www.hanen.org		**Grade/Age Range:** 2- through 4-year-old (toddlers and preschoolers) classrooms **Administration Interval:** As frequently as desired
Material, Training, and Scoring Costs: Forms (50 with user's guide): $35.00		**Personnel and Training Requirements** **Credentials Required for Use:** Level 2+ (certification beyond bachelor's like a master's) **Personnel for Administration:** Highly trained individual **Training for Administration:** Self-training (<1 hour) Although not required, the publisher recommends that observers undergo training on the scale as a part of the *Learning Language and Loving It* program training to increase scoring accuracy and reliability.[1]
Languages: English		**Alternate Forms:** No
Representativeness of Norming Sample: No norming sample		**Summary** **Initial Material Cost:** 1 (<$100) **Time to Administer:** 5 to 10 minutes **Ease of Administration and Scoring:** 3 (administered and scored by a highly trained individual) **Reliability:** 3 (all at or above 0.70) **Predictive Validity:** Not available **Construct/Concurrent Validity:** Available **Norming Sample Characteristics:** 1 (none described)

NARRATIVE

Description: The Teacher Interaction and Language Rating Scale is a classroom observation tool that evaluates teachers' and caregivers' interactions with small groups of 2- through 4-year-olds in a group setting. It focuses on interactions related to techniques that are expected to increase children's language acquisition. The observer watches a videotape of teacher-child interactions and records ratings for 11 items (i.e., techniques) on a paper form. Although developers do not specify an observation period, two of their studies included 10-minute videotaped sessions (Girolametto and Weitzman 2002; Girolametto et al. 2000), and the publisher's web site describes 5-minute videotaped sessions for training. The scale evaluates teachers' use of three strategies resulting in subscale scores: Child-Oriented (utterances that follow the child's lead in terms of topic and activity; four items), Interaction-Promoting (utterances encouraging children to engage in extended conversational turns; three items), and Language-Modelling (utterances that expand or extend the semantic content of children's communicative attempts; four items). One Language-Modelling item called Imitate is used only with toddlers because it involves children who are preverbal or at the one-word stage of language production. The paper form includes a description of each item, which may include several skills. Observers score items on a 7-point scale indicating the frequency (from almost never to consistently) with which teachers use a technique.

Other Languages: None.

Uses of Information: Observers may use Teacher Interaction and Language Scale ratings to document teachers' status and progress with techniques that foster students' language development. Developers also note that results from the scale may be used to set program goals. The scale was designed to evaluate teachers' interaction with students before and after being trained on the *Learning Language and Loving It* program, which covers the 11 techniques.

Methods of Scoring: To address research questions, developers using the scale have calculated mean ratings by subscale (Child-Oriented, Interaction-Promoting, and Language-Modelling) and item. Observers may calculate the total score by adding the ratings, each on a 7-point scale, for the 11 techniques. Anchors include 1 (almost never), 3 (sometimes), 5 (frequently), and 7 (consistently). An observer may select the rarely used response option N/A if the technique is not appropriate for the classroom activity or the child's age or the teacher does not need to demonstrate skill for the technique.

Interpretability: The Teacher Interaction and Language Rating Scale form includes guidelines for observers' use in interpreting individual item ratings. Item ratings of 1 through 4 indicate that interactive techniques need improvement or fine-tuning and may be pinpointed as goals for improving future interactions with students. Ratings of 5 through 7 indicate that teachers' interactive techniques achieve expectations and should not be pinpointed as goals for improvement.

Reliability:
(1) Internal consistency reliability: No information available.
(2) Test-retest reliability: No information available.
(3) Alternate form reliability: Not applicable.

(4) Inter-rater reliability: Percentages of agreement for ratings within one scale point ranged from 82 to 100 percent by item (Girolametto and Weitzman 2002; Girolametto et al. 2000) and 89 percent overall (Girolametto and Weitzman 2002). Ratings were based on eight videotaped teacher-child groups with a two-item scale and six videotaped teacher-child groups with a nine-item scale. In addition, the authors examined intra-reliability for observers by examining changes in a single observer's ratings over time (rater drift) of eight videotaped teacher-child groups. Intra-rater reliability was 100 percent based on the two-item scale with ratings one month apart (Girolametto et al. 2000) and the nine-item scale with ratings two weeks apart (Girolametto and Weitzman 2002).

Validity Evidence:
The scale evaluates 11 behaviors derived from *Learning Language and Loving It*, a program for early childhood educators that includes three strategies to improve children's language acquisition (Child-Oriented, Interaction-Promoting, and Language-Modelling). The strategies are based on a social interactionist perspective of language development in which the rate of children's language acquisition varies according to differences in language input from teachers and caregivers (Bohannon and Bonvillian 1997).

Construct/Concurrent validity: Girolametto and Weitzman (2002) showed with repeated measures multivariate analysis of variance (MANOVA) tests that item ratings differed according to the observed context. For example, results from the three Interaction-Promoting items showed that teachers facilitating a play dough activity had higher ratings than those leading a book-reading activity.

Pearson product moment correlation coefficients between the mean Child-Oriented subscale ratings and three measures of preschoolers' language productivity were 0.64 for the number of utterances, 0.49 for the number of different words, and 0.62 for the number of multiword utterances (Girolametto and Weitzman 2002). Coefficients between the mean ratings for the Interaction-Promoting and Language-Modelling subscales and the three language productivity measures were 0.65, 0.54, and 0.62 and 0.51, 0.41, and 0.48, respectively. The sample included 56 preschoolers and toddlers (data on toddlers not included here), with an even split by gender, and 26 child care providers from licensed, nonprofit child care centers in Toronto.

In another study, Pearson product moment correlation coefficients between the reverse-scored Encourage Turn Taking mean item rating (previous item version used) and number of utterances, different words, and multiword combinations were -0.60, -0.50, and -0.55, respectively, demonstrating that teachers' restriction of verbal turn-taking was associated with restricted and less complex language use by children (Girolametto et al. 2000). Correlations were not significant between that item and the longest utterance, another language productivity measure. The Follow the Children's Lead mean item rating (previous item version used) was not significantly correlated with any of the language productivity measures. The sample included 80 toddlers and preschoolers (half of whom were female) and 20 early childhood educators from licensed, nonprofit child care centers in Toronto.

Authors also examined mean ratings for Child-Oriented, Interaction-Promoting, and Language-Modelling subscales by child age, testing the hypothesis that certain subscales would have higher ratings depending on the age of children and the observation context (Girolametto and Weitzman

2002). Repeated measures MANOVA tests showed no significant main effects of age with any of the subscales; however, when authors tested interactions between age and context, the Language-Modelling subscale ratings differed according to whether children were toddlers or preschoolers. In particular, toddler teachers had higher ratings for Use a Variety of Labels (i.e., vocabulary), and preschool teachers received higher ratings for Extend (i.e., teachers provide information related to children's topics).

Predictive validity: No information available.

Bias Analysis: No information available.

Training Support: While not requiring training for use of the Teacher Interaction and Language Rating Scale, the publisher provides a three-day workshop (at a cost of approximately $700) on the *Learning Language and Loving It* program for speech-language pathologists/therapists, early childhood education consultants, literacy specialists, and professors and instructors who will then train early childhood educators on the program. As part of the training, observers learn to evaluate videotaped teacher-child interactions to assess whether teachers apply the 11 techniques and to make ratings on the scale. The publisher does not offer exclusive training on the scale. The publisher's web site does not describe inter-rater reliability training, but two studies discussed the training of observers to achieve a level of 85 percent agreement within one scale point with the developer for all ratings (Girolametto and Weitzman 2002; Girolametto et al. 2000).

Adaptations/Special Instructions for Individuals with Disabilities: For children who use sign language, a picture communication system or alternative communication, observers are instructed to interpret "gestures, sounds, and words" as these forms of communication.

Alternate Forms: Not applicable.

Previous Version: The previous version of the Teacher Interaction and Language Rating Scale included 14 items and assessed the quality, completeness, and consistency with which teachers carried out techniques described in each item (Girolametto et al. 1999). The 7-point response ratings ranged from "inadequate" to "excellent."

NCEE or REL Study Use:[2] Accelerating language development in kindergarten through Kindergarten PAVEd for Success

[1]J. Greenberg, personal communication, February 10, 2009.

[2] See Table F.1 for web address.

References:

Bohannon, John, and John Bonvillian. "Theoretical Approaches to Language Acquisition." In *The Development of Language: Fourth Edition*, edited by Jean Berko Gleason. Needham Heights, MA: Allyn and Bacon, 1997.

Girolametto, Luigi, Elaine Weitzman, and Riet van Lieshout. "The Teacher Interaction and Language Rating Scale." Unpublished manuscript. Toronto, ON: University of Toronto, 1999.

Girolametto, Luigi, Elaine Weitzman, and Janice Greenberg. *Teacher Interaction and Language Rating Scale*. Toronto, ON: The Hanen Centre, 2000.

Girolametto, Luigi, Elaine Weitzman, and Riet van Lieshout. "Directiveness in Teachers' Language Input to Toddlers and Preschoolers in Day Care." *Journal of Speech, Language, and Hearing Research*, vol. 43, 2000, pp. 1101-1114.

Pepper, Jan, and Elaine Weitzman. *It Takes Two to Talk: A Practical Guide for Parents of Children with Language Delays: Third Edition*. Toronto, ON: The Hanen Centre, 2004.

Weitzman, Elaine, and Janice Greenberg. *Learning Language and Loving It: A Guide to Promoting Children's Social, Language, and Literacy Development in Early Childhood Settings*. Toronto, ON: The Hanen Centre, 2002.

TABLE D.1

NCEE OR REL RECENTLY DEVELOPED CLASSROOM PRACTICES AND SETTINGS MEASURES SUMMARY

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range [b]	Reliability	Validity
Comprehensive Classroom Practices					
Classroom Characteristics (CC) form	Math Curricula (2)	This measure is a classroom observation of classroom quality and school engagement. On a scale from 1 (not at all) to 4 (extremely characteristic), assessors rate the degree to which a statement is characteristic of the observed class. The statements focus on behavior management, use of instructional time, the classroom's social environment (e.g., "Teachers and students have a warm, positive relationship"), instructional delivery (e.g., monitoring of instruction), and student engagement and involvement in the classroom. Observers must pass a reliability test, reaching 80 percent agreement on the CC ratings. The study enrolled 110 schools in 12 school districts.	Grades 1–3	Not available	The CC form was piloted with videotapes of classrooms and in live classrooms. The protocols were revised after the study's advisory panel and the Institute of Education Sciences (IES) provided feedback.
Teacher questionnaire of attitudes and behaviors	Formative Assessment (REL-Midwest) (3)	This measure is a teacher self-report of comprehensive instructional practices. Topics covered in the questionnaire include instructional strategies, teacher collaboration, knowledge of use of data to guide instruction, strategies for differentiating instruction, and attitudes toward professional development activities. The questionnaire requires 45 minutes for completion and is administered online. Approximately 168 teachers in 42 schools will complete the questionnaire.	Grades 4–5	Not available	The questionnaire uses items adapted from the Study of Instructional Improvement (University of Michigan) and the Surveys of Enacted Curriculum (Wisconsin Center for Education Research and Mathematica Policy Research, Inc.).

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.

TABLE D.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range [b]	Reliability	Validity
Teacher questionnaire of classroom quality and instructional practices	Different Routes to Certification (2)	This measure is a teacher self-report of instructional practices (specifically for reading and mathematics), classroom quality (classroom environment), and motivation for teaching. Other questions in the questionnaire focus on teachers' background, the teacher preparation program, support activities, and opinions about the school. The items on instructional practices for reading and mathematics ask the teacher to report the number of minutes the students received instruction and devoted to homework, the time spent in different instructional modes for reading and mathematics (e.g., teacher-directed whole-class activities), and the frequency (6-point scale from never to daily) with which students participated in various reading and mathematics activities, such as performing plays and skits or playing mathematics-related games. Classroom quality questions include whether student misbehavior hindered the teacher's ability to teach effectively. Motivation for teaching includes commitment to the school and to teaching (e.g., the number of years the teacher plans to continue teaching). The study enrolled 68 schools and approximately 180 teachers.	Kindergarten through grade 5	Not available	Not available
Teacher questionnaire of educational practices	School Improvement Intervention (REL-Central) (3)	This measure is a teacher self-report of instructional practices (comprehensive in nature), school engagement, and school climate. The questions address data-based decision making, shared leadership, purposeful community, and effective school practices. The measure's five sections include a total of approximately 125 items. The School Environment section contains questions about parental involvement (7	Grades 3–5	Not available	The items on purposeful community in the Professional Community section came from R. Goddard (2000). The rest of the items are from McREL (2005).

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.

TABLE D.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range [b]	Reliability	Validity
		items), academic press (8 items), safe and orderly climate (7 items), and assessment and monitoring (8 items).The Professional Community section includes items on collaboration/deprivatization (8 items), professional development (8 items), support for teacher influence (8 items), and purposeful community (12 items). The Leadership section contains items on shared mission and goals (8 items), instructional guidance (8 items), and organizational change (10 items). The Instruction section contains questions on individualization (9 items), structure (8 items; e.g., students "independently manage their classwork," or "receive written or verbal feedback on their progress"), and opportunity to learn (9 items). The last section asks questions about the teacher's background. Typically, responses range from "strongly agree" to "strongly disagree" on a 5-point scale, except for the items on collaboration/deprivatization, professional development, structure, and opportunity to learn, which are rated on a 5-point scale from "great extent" to "not at all." The questionnaire takes approximately 25 minutes for completion and is administered online. The study plans to enroll 52 schools, and 1,040 teachers will complete the questionnaire.			

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.

TABLE D.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range [b]	Reliability	Validity
Reading Practices					
Classroom observations of instructional quality	Adolescent Literacy Across the Curriculum (REL-Midwest) (3)	This measure is a classroom observation of instructional practices in reading as well as of student reading practices. Observers observe an entire day of instruction at the school, following randomly selected student schedules. The study plans to complete classroom observations in approximately 40 schools.	Grades 9–12	Not available	This measure is an adapted version of a previously validated classroom observation tool developed by Learning Point Associates and used in another IES study.
Classroom observation of literacy teaching practices	Accelerating language development (REL Southeast) (3)	This measure is a classroom observation of instructional practices pertaining to vocabulary and literacy. The classroom observation tool is a time-sampling measure with three parts. First, every five minutes the observer completes a 30-second Snapshot of Classroom form that records the occurrence of 10 aspects of literacy content (e.g., comprehension support), noting instructional delivery (e.g., presenting or interacting with student(s)) as well as specific context (e.g., lesson in literacy content, class meeting). Second, during the four minutes between coding intervals, the observer completes the Teacher Talk Observation Profile (TOP), which records the frequency of 12 teacher behaviors in two categories: teacher talk and vocabulary. The observer notes whether the behavior occurred in teacher-led instruction or in teacher-student interactions. For teacher-led instruction (both teacher talk and vocabulary) and for teacher-student interactions about vocabulary, the assessor marks an X for each occurrence of the behavior. For teacher talk using teacher-student interactions, the assessor codes each occurrence of teacher talk with a S (statement) or Q (question). Third, the observer documents, on a Teacher Read	Kindergarten and grade 1	Not available	Not available

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.

TABLE D.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range[b]	Reliability	Validity
		Aloud form, the occurrence of teachers' use of reading strategies, types of talk, and vocabulary instruction before, during, and after reading aloud to students. Observers note whether students are involved and to whom the teacher is reading (e.g., whole class or individual) and then rate the conversations (e.g., some back and forth occur). The classroom observation occurs for the entire length of time devoted to literacy instruction during the school day. The sample size is targeted for 60 to 80 schools with a sample of approximately 160 teachers.			
Expository Reading Comprehension Classroom Observation (ERCCO)	Reading Comprehension (1, 2, 3)	This measure is a classroom observation of instructional practices (specifically reading). It captures the level of implementation of reading comprehension interventions and teachers' use of comprehension teaching strategies. During 10-minute observation intervals throughout the instructional period, the observer records data on the frequency of behaviors indicative of high-quality comprehension (8 behaviors) and vocabulary (6 behaviors) instruction (e.g., teacher "asks students to justify or elaborate their responses"). The observer records the number of times the behavior happened as teacher modeling, teacher explaining, or student practice. The observer also notes other behaviors and classroom characteristics that are not outcomes, including student grouping patterns, type of text (e.g., independent silent reading or teacher reads aloud), overall instructional quality, classroom management, teacher responsiveness to students, and levels of student engagement at the class level.	Grade 5	From the Evaluation of Reading Comprehension Interventions: Item Response Theory (IRT) reliability for the three measures derived from the ERCCO: Traditional Interaction: 0.70 Reading Strategy Guidance: 0.72 Classroom Management: 0.83 Inter-rater reliability: Traditional Interaction: 0.97-0.98 Pearson correlation between	The Evaluation of Reading Comprehension Interventions study examined criterion validity of the ERRCO by examining correlations between its scales and student achievement scores (composite test scores; Group Reading Assessment and Diagnostic Evaluation [GRADE]; Social Studies Reading Comprehension Assessment; Science Reading Comprehension Assessment). Traditional Interaction: No statistically significant correlations with achievement Reading Strategy Guidance: Statistically significant correlations (r = 0.07 to 0.09) Classroom Management:

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.

TABLE D.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range [b]	Reliability	Validity
		The Evaluation of Reading Comprehension Interventions study selected 89 schools and 268 teachers for ERCCO observations, and the Assessing the Impact of Collaborative Strategic Reading study planned ERCCO observations in approximately 40 classrooms.		scale scores; 86 percent agreement at item level Reading Strategy Guidance: 0.97 Pearson correlation between scale scores; 90 percent agreement at item level Classroom Management: 0.94 Pearson correlation between scale scores	Statistically significant correlations (r = 0.09 to 0.13)
Instructional Practice in Reading Inventory (IPRI)	Reading First (1)	This measure is a classroom observation of instructional practices with a focus on five dimensions of reading instruction: (1) phonemic awareness, (2) decoding/phonics, (3) fluency, (4) vocabulary, and (5) comprehension. The instrument also collects other information about instruction, including oral reading by students (if the teacher has not clearly indicated the instructional purpose of the oral reading), oral reading by teacher alone, silent reading, spelling, written expression, other language arts, assessment, non-literacy instruction, non-instruction, academic management, transitions between activities, and interruptions to instruction to manage student behavior. The observer records the occurrence of the specified teacher instructional behaviors during continuous 3-minute observation intervals. The observer uses a new IPRI form for each observation	Grades 1–2	Inter-rater reliability: Spring 2005: 88 percent agreement Fall 2005: 90 percent agreement	Sources contributing to the development of the IPRI include (1) scientifically based research on effective elementary grade reading instruction; (2) reviews of existing classroom observation tools of instructional practices and content, especially those tapping language, literacy, and reading; and (3) research on the development of classroom observation measures.

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.

[b]Study sample was used to determine grade/age range of measure.

TABLE D.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range[b]	Reliability	Validity
		interval during the entire scheduled observation period. Most observation periods are 90 minutes or more, depending on the schools' defined reading blocks. IPRI observations were completed in 8,670 classrooms in 248 schools.			
Lexical diversity	Accelerating language development (REL Southeast) (3)	This measure is a classroom observation of classroom quality, specifically the teacher's lexical diversity directed to students in the classroom. Lexical diversity is the number of unique words relative to the total number of words spoken. During the classroom observation visit, the observer tape records 20 minutes of the teacher's instruction during a small-group activity. The audiotape is then analyzed by using a language analysis program. The sample size is targeted for between 60 and 80 schools with a sample of approximately 160 teachers.	Kindergarten and grade 1	Not available	Not available
Teacher questionnaire on reading instructional strategies	Reading First (1)	This measure is a teacher self-report of instructional strategies (reading). The questionnaire taps four areas assessed as outcomes: (1) professional development, (2) the amount of reading instruction, (3) supports for struggling readers, and (4) use of assessments. The professional development area consists of items about the amount of professional development in reading received by teachers (e.g., hours attending workshops or conferences on reading), whether the teacher received professional development in the five essential components of reading instruction (phonemic awareness, phonics, vocabulary, fluency, and comprehension), and whether the teacher received help from a reading	Grades 1–3	Cronbach's alpha: Amount of professional development in reading received by teachers: 0.22 Teacher receipt of professional development in the five essential components of reading instruction: 0.86 Teacher receipt of coaching: Not applicable because it is a single	Not available

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.

TABLE D.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range [b]	Reliability	Validity
		coach. The amount of reading instruction is derived from the number of minutes devoted to reading instruction per day as reported by teachers. Supports for struggling readers (four items) examine the receipt of extra classroom practice for struggling readers over the past month for several topics (phonemic awareness, phonics, fluency, and comprehension). For the use of assessments (three items), teachers report on the use of test results for a specific purpose (e.g., to organize instructional groups) and the degree to which assessment is part of their reading instruction (central to, small part of, or not). Other questionnaire questions not examined as outcomes of the study include teachers' instructional strategies, materials used, other types of assessments used, other supports for struggling readers, and types of professional development. From the 248 schools enrolled in the study, approximately 2,000 teachers completed the questionnaire.		dichotomous outcome variable Minutes spent on reading instruction each day: 0.99 Provision of extra classroom practice for struggling readers: 0.77 Use of assessments to inform classroom practice: 0.60	
Mathematics Practices					
Algebra I Quality Assessment (AQA)	Hybrid Algebra I (REL-Appalachia) (3)	This measure is a classroom observation of instructional practices pertaining to Algebra I instruction. It draws on the Kentucky Virtual High School model of instruction and is designed to gather more detailed information about Algebra I instructional practices than the Algebra I teacher questionnaire (see entry in this table). The observation is completed during Algebra I classes and lasts approximately 60 minutes. Observers will complete approximately 300 classroom observations in about 60 schools.	Grade 9	Not available	Not available

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.

TABLE D.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range [b]	Reliability	Validity
Algebra I teacher questionnaire	Hybrid Algebra I (REL-Appalachia) (3)	This measure is a teacher self-report of instructional practices (specifically mathematics) and school engagement. The questionnaire includes items on the practices teachers use when teaching Algebra I (e.g., "Using number lines, graphs, or diagrams to explain Algebra"), the activities of students in the class (e.g., how often students work in groups), and student engagement at the class level (1 item for agreement—"Student interest and engagement is high when I use the Hybrid Algebra I Approach"). Most of the questions on instructional practices are answered on a 5-point frequency scale (never to extensively). In addition, the questionnaire includes items (not used as outcomes) on teachers' attitudes toward the Algebra I curriculum that they use, the activities available for teachers in the school, and interactions with other teachers. This paper-based questionnaire requires about 10 minutes for completion. The study has a target sample size of 60 high schools and 120 teachers.	Grade 9	Not available	Nine Algebra I teachers in Kentucky schools pre-tested the teacher questionnaire. The teachers did not provide any suggestions for revisions.
Classroom observation of math practices	Professional Development Strategies in Math (3)	This measure is a classroom observation of instructional practices (specifically mathematics) and school engagement. It records six categories of behavior: Explanation/Instruction (8 items), Questioning/Feedback (6 items), Lesson Structure (3 items), Representations (12 items), Delivery (7 items), and Student Engagement (5 items; at the classroom level). The observations occur during one class session and last from 45 to 90 minutes. Approximately 504 classroom observations	Grade 7	Not available	The measure was based on tools developed for other studies; including the TIMMS Video study (Hiebert et al. 2003), the Cognitively Guided Instruction Study (Carpenter et al. 1989), and the QUASAR project (Silver and Stein 1996).

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.

TABLE D.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range [b]	Reliability	Validity
		in 128 schools will be completed.			
Observation of Math Instruction (OMI) form	Math Curricula (2)	This measure is a classroom observation of mathematics instructional practice and classroom quality (specifically, classroom environment). Assessors complete the OMI form during mathematics classes by using interactive coding, checking off clearly defined behaviors as they occur. Observers note instructional time, teacher-initiated instructional behaviors, teacher feedback to students, use of metacognitive strategies, evidence of instructional behaviors (e.g., whether the teacher states the lesson objective at the beginning of class), use of representations, the proportion of students in the class involved in different types of activities (e.g., played math games), focus of mathematics practice (e.g., "Number of practice problems focused on today's objective"), materials used by students, problem-solving approaches, and activity setting. For many of the activities recorded on the OMI form, the assessor not only notes whether a behavior occurs but also indicates its frequency. The observation covers all mathematics instruction that occurs that day. The assessors must pass a reliability test, reaching 80 percent agreement on coding the categories on the OMI form and ratings. The study enrolled 110 schools in 12 school districts.	Grades 1–3	Not available	Development of the OMI form was based on a review of the literature (mathematics instruction and classroom observations) and analysis of the curricula included in the study. It was piloted with videotapes of classrooms and in live classrooms. The protocols were revised after the study's advisory panel and IES provided feedback.
School Engagement or Climate					
Student and parent questionnaires of school climate	DC Opportunity Scholarship (1)	This measure consists of student and parent self-reports of school climate (specifically of safety) and satisfaction with the school. For school safety, the parent rates the perceived seriousness of 10 problems at the student's school, such as fighting or cheating (ratings	Grades 4–12	Cronbach's alpha: Parent satisfaction scale: 0.93 Student satisfaction scale: 0.85 No information	Not available

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.

[b]Study sample was used to determine grade/age range of measure.

TABLE D.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range[b]	Reliability	Validity
		on a 3-point scale from not serious to very serious). On the student questionnaire, students report the number of times--"never" to "three or more"--that they experienced eight events in the past year, including theft or assault. The items on parent satisfaction ask parents to grade their child's school (from F to A) and rate 12 items about the school (e.g., class sizes, location, or academic quality) on a 4-point scale from "very dissatisfied" to "very satisfied." Students also give their school a grade for overall satisfaction and rate their agreement on 17 items about their school (e.g., "There is a lot of learning at the school" and "My teachers are fair") on a 4-point scale from "disagree strongly" to "agree strongly." The questionnaires each take approximately 15 minutes for completion. The study plans to have approximately 2,308 students and 2,308 parents complete the questionnaires.		available for parent or student reports of school safety	

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.

TABLE D.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range [b]	Reliability	Validity
Student questionnaire of behaviors and violence	School-Based Violence Prevention (3, 4)	This measure is a student self-report of school climate, specifically opinions about safety at school. The questionnaire also collects reports on social-emotional behaviors (see Table B.1). The full questionnaire consists of 85 items that include background questions and items focused on getting along with people at school and the student's feelings and attitudes. The questions about safety at school ask students to rate how frequently--from "never" to "often"--they are afraid that another student will attack, harm, or bully them. The full questionnaire takes approximately 45 minutes for completion and is a self-administered paper questionnaire. Approximately 36,920 students in 40 middle schools will complete the student questionnaire.	Grades 6–8	No information available for student safety concerns	The items on student safety concerns at school were adapted from the School Crime Supplement to the National Crime Victimization Survey.
Teacher questionnaire of school climate	Lessons in Character Education (REL-West) (3)	This measure is a teacher self-report of school climate, particularly school expectations (5 items; e.g., "The students in this school are expected to tell the truth"), parent and staff relations (7 items), and staff culture of belonging (10 items). The questionnaire also asks teachers about classroom practices, their backgrounds, and their professional development activities. Approximately 750 teachers from 50 schools will complete the questionnaire	Grades 2–5	Cronbach's alpha: School expectations: 0.94 Parent and staff relations: 0.88 Staff culture of belonging: 0.91	Items were derived from *Evaluation resource guide: Tools and strategies for evaluating a character education program (*Characterplus 2002).

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.

TABLE D.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range [b]	Reliability	Validity
Teacher questionnaire on safety and victimization	School-Based Violence Prevention (3, 4)	This measure is a teacher self-report of school climate, specifically sense of safety and victimization. Teachers base their responses on their perceptions over the last 30 days about safety concerns (six items) and their victimization involving any verbal or physical threats or attacks (three items). Additional items in the questionnaire include questions about the school, the frequency of violent events between students that teachers witnessed, and teachers' techniques for dealing with aggressors and victims. The questionnaire takes about 30 minutes for completion. A sample of 24 teachers at each of the 40 participating schools will complete the questionnaire, resulting in a sample of 960 teachers.	Grades 6–8	Cronbach's alpha for each subscale from the current study: Teacher safety concerns: 0.89 Teacher victimization: 0.68	The items on teacher safety concerns were adapted from the School Crime Supplement to the National Crime Victimization Survey. The items on teacher victimization were adapted from the Schools and Staffing Survey (SASS).
Other/Multidomain					
Student questionnaire of economic interest and attitudes	Problem-Based Economics (REL-West) (3)	This measure collects student self-reports of instructional practices in economics as well as student-level outcomes on approaches to learning/motivation (see Table B.1). On a 5-point scale from "not at all" to "very much", students rate the degree to which 13 class activities helped them learn economics (e.g., reading the textbook). In addition to the outcomes noted above, the questionnaire captures students' attitudes toward school, their school behaviors, and their problem-solving skills. The study will administer questionnaires to 4,800 students in 40 schools.	Grade 12	A previous version was reported to have a Cronbach's alpha of 0.80; no other information available	The questionnaire uses several items from the Student Assessment of Learning Gains, developed by the Wisconsin Center for Educational Research. In addition, economists outside of the study team reviewed the questionnaires during development.

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.

TABLE D.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range [b]	Reliability	Validity
Teacher questionnaire of instructional practices	Alabama Math, Science, and Technology Initiative- AMSTI (REL-Southeast) (3)	This measure is a teacher self-report of instructional practices (mathematics and science). Teachers are asked to recall their instruction during the previous two weeks and report the number of minutes spent teaching mathematics and science. Teachers indicate the types of assessments used in their classrooms and report the number of minutes that students engaged in Inquiry-Based Instruction, hands-on activities, and higher-order thinking for both mathematics and science. Over two years, teachers complete four versions of the questionnaire, which are web-based. The questionnaire requires approximately 20 minutes for completion. The study expects approximately 324 teachers in 40 schools to complete the questionnaire.	Grades 4–8	Not available	Items were modified from the following sources: (1) Integrated Studies of Educational Technology Teacher Survey; (2) the National Educational Technology Trends Study: Teacher Survey; and (3) the Empirical Education Item Bank.
Teacher questionnaire of instructional practices and self-efficacy	Principles-Based Professional Development (REL-Pacific) (3)	This measure is a teacher self-report about instructional practices (comprehensive and language arts) and motivation for teaching (self-efficacy). The questionnaire includes approximately 20 questions about teachers' frequency of instructional practices, including items related to teaching students who are English language learners. The questionnaire also asks teachers to rate their agreement on statements about student learning (5-point scale from strongly disagree to strongly agree) and rate the extent to which 11 factors pose a challenge in teaching to the language arts standard (4-point scale from not at all to a great deal). These items address areas such as students' ability, teachers' beliefs (e.g., "The use of native language at home can impede learning a second language"), and teachers' self-	Grades 4–5	Not available	Not available

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.

TABLE D.1 (*continued)*

Measure	NCEE or REL Study (source) [a]	Description	Grade/Age Range [b]	Reliability	Validity
		efficacy (e.g., whether teachers' limited knowledge of the content area impedes their ability). The questionnaire also collects information on teachers' beliefs about the effectiveness of intervention activities. The questionnaire is administered online. Approximately 270 teachers from 50 schools are expected to complete the questionnaire.			
Teacher questionnaire of practices and economic attitudes	Problem-Based Economics (REL-West) (3)	This measure is a teacher questionnaire of instructional practices (in economics) and motivation for teaching (confidence and enthusiasm to teach economics). On a series of 5-point scales, the questionnaire measures teachers' use of constructivist-oriented teaching methods. Teachers rate (1) how often they use 12 methods to teach economics (e.g., have students use the internet to get information) from "never" to "almost every day"; (2) their economic content knowledge (from poor to excellent); (3) their confidence in teaching key economics concepts (11 items, from not very confident to totally confident); and (4) their enthusiasm for teaching economics in the future (4 yes/no items; e.g., "In the future, I am willing to teach Economics if assigned"). Other questions (not considered outcomes) in the questionnaire include barriers to teaching economics, such as student interest, and whether teachers received professional development. The questionnaires require 5 to 10 minutes for completion. The study will administer questionnaires to approximately 120 teachers in 40 schools.	Grade 12	Cronbach's alpha for original scales: 0.84 (Ravitz and Mergendoller 2005, 6-item index on economic content knowledge and confidence teaching economics) 0.90 (Ravitz et al. 2000, 7- item index on instructional practices) No other information available	Items come from earlier work by the Buck Institute for Education. Items on teacher pedagogy come from *Teaching, Learning and Computing* (1998), and items on economic content knowledge and confidence in teaching economics were drawn from Ravitz and Mergendoller (2005). Items on teacher instructional practices were drawn from Ravitz et al. (2000).

[a]Source codes are 1 = study results or methodology report, 2 = other documentation (codebook, study design report), 3 = OMB clearance package, and 4 = personal communication.
[b]Study sample was used to determine grade/age range of measure.

RECENTLY DEVELOPED CLASSROOM MEASURE REPORT REFERENCES

Agodini, Roberto, John Deke, Sally Atkins-Burnett, Barbara Harris, and Robert Murphy. "Design for the Evaluation of Early Elementary School Mathematics Curricula." Submitted to the Institute of Education Sciences, U.S. Department of Education. Princeton, NJ: Mathematica Policy Research, January 2008.

Agodini, Roberto, Barbara Harris, Sally Atkins-Burnett, Sheila Heaviside, Timothy Novak, and Robert Murphy. "Achievement Effects of Four Early Elementary School Math Curricula: Findings from First Graders in 39 Schools." (NCEE 2009-4052). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, February 2009.

Constantine, Jill, Daniel Player, Tim Silva, Kristin Hallgren, Mary Grider, and John Deke. "An Evaluation of Teachers Trained through Different Routes to Certification, Final Report." (NCEE 2009-4043). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, February 2009.

Decker, Paul, John Deke, Amy Johnson, Daniel Mayer, John Mullens, and Peter Schochet. "The Evaluation of Teacher Preparation Models: Design Report." Submitted to the National Center for Education Evaluation, Institute of Education Sciences, U.S. Department of Education. Princeton, NJ: Mathematica Policy Research, October 2005.

Dimino, Joseph, Mary J. Taylor, Madhavi Jayanthi, Rebecca Newman-Gonchar, Jan Dole, Lauren Liang, Meaghan Edmonds, and Sharon Vaughn. "Evaluation of Reading Comprehension Interventions: Expository Reading Comprehension Classroom Observation Instrument Observer's Guide." Austin, TX: RG Research Group, University of Utah, University of Texas at Austin, 2006.

Gamse, Beth, Howard Bloom, James Kemple, Robin T. Jacob, Beth Boulay, Laurie Bozzi, Linda Caswell, Megan Horst, W. C. Smith, Robert St. Pierre, Fatih Unlu, Corinne Herlihy, and Pei Zhu. "Reading First Impact Study: Interim Report." (NCEE 2008-4016). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, April 2008.

Gamse, Beth C., Robin T. Jacob, Megan Horst, Beth Boulay, and Fatih Unlu. "Reading First Impact Study Final Report." (NCEE 2009-4038). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, November 2008.

James-Burdumy, Suzanne, David Myers, John Deke, Wendy Mansfield, Russell Gersten, Joseph Dimino, Jan Dole, Lauren Liang, Sharon Vaughn, and Meaghan Edmonds. "The National Evaluation of Reading Comprehension Interventions: Design Report." Submitted to the Institute of Education Sciences, U.S. Department of Education. Princeton, NJ: Mathematica Policy Research, May 2006.

James-Burdumy, Susanne, Wendy Mansfield, John Deke, Nancy Carey, Julieta Lugo-Gil, Alan Hershey, Aaron Douglas, Russell Gersten, Rebecca Newman-Gonchar, Joseph Dimino, and Bonnie Faddis. "Effectiveness of Selected Supplemental Reading Comprehension Interventions: Impacts on a First Cohort of Fifth-Grade Students." (NCEE 2009-4032). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

Wolf, Patrick, Babette Gutmann, Michael Puma, Brian Kisida, Lou Rizzo, and Nada Eissa. "Evaluation of the DC Opportunity Scholarship Program: Impacts After Three Years." (NCEE 2009-4050). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, March 2009.

Wolf, Patrick, Babette Gutmann, Michael Puma, Lou Rizzo, Nada Eissa, and Marsha Silverberg. "Evaluation of the DC Opportunity Scholarship Program: Impacts After One Year." Submitted to the Institute of Education Sciences, U.S. Department of Education. Washington, DC: U.S. Government Printing Office, June 2007.

APPENDIX E

GLOSSARY OF TERMS

GLOSSARY OF TERMS[1]

Accommodation. A modification in materials, administration procedures, or other change in a test to accommodate the needs of students with disabilities or limited English proficiency. The goal is to allow these individuals to be included in the assessment without altering what the assessment measures or the comparability of scores. Those administering the assessment provide the accommodations, which may include use of assistive devices, additional time, or adaptations in the language of administration (for example, the use or acceptance of non–English responses).

Adaptive. An assessment that routes students through items based on their answers to specific items. If students answer the items correctly, they move to a more challenging sequence of items. If not, students move to easier items. Routing reduces overall assessment time and burden on the student and allows for a more refined assessment of skills. Computer administration can provide fully adaptive assessments. Individual administration can provide adaptive assessment by way of basal or ceiling rules (see Basal and Ceiling). Two-stage (or multistage) adaptive tests may be group- or individually-administered. In a two-stage adaptive test, the first stage is a shorter assessment that routes students into a second-stage form targeted to their ability. If a student is not required to respond to all items, the assessment is termed adaptive.

Adjusted validity/reliability coefficient. A validity or reliability coefficient that has been adjusted to offset effects of differences in score variability, criterion variability, or unreliability of test and/or criterion. When variance is limited in one of the measures, correlations are lower. Similarly when there is more measurement error, correlations between measures will be lower. The adjusted coefficient may be referred to as corrected for range or corrected for restricted range. (See Restriction of range or variability.)

Age equivalent score. Score based on the median raw score of students at that age in the norming sample. The age-equivalent score corresponding to a student's raw score provides information on the student's level of performance in terms of the age at which that level of performance on that assessment could be expected, based on the norming sample. An age-equivalent score can be misleading in that the content on the assessment may not include content appropriate for all age levels for which an age-equivalent score is estimated. Students of the same age may be in different grades with different opportunities to learn. (See Grade equivalent score, Raw score, Norming sample.)

Alignment. The process to ensure that specific aspects of the educational process (for example, assessment, instruction, curriculum) are in line with each other, particularly the development of standards for content and performance and the implementation/availability of classroom activities and materials to support those standards. Assessments should also be

[1] The definitions in this glossary were developed using a variety of sources, a list of which can be found at the end of this appendix.

designed to align with set standards to ensure accurate measurement of appropriate content and desired outcome. The goal is for the curriculum to teach what the students are expected to learn and for the assessments to measure the extent to which students have learned the relevant items. In the context of a study of an educational intervention, alignment of the measurement approach with the intervention means that the measures assess areas expected to be affected by the intervention (for example, a mathematics intervention should include measures of the quality of mathematics teaching and of students' mathematics achievement).

Alternate form. Two or more versions of one test that are considered interchangeable because they purportedly measure the same constructs in the same ways. The alternate forms are intended for the same purpose and administered with the same directions. Alternate forms is a generic term used to describe tests in any of three categories: (1) parallel forms have equal raw score means, standard deviations, error structures, and correlations with other measures for any given population; (2) equivalent forms do not demonstrate statistical similarity, but the differences in raw score statistics are compensated for in conversion to derived scores or in the forms' norm tables; or (3) comparable forms are similar in content but have no demonstrated statistical similarity. (See Alternate form reliability, Parallel forms.)

Alternate form reliability. Publishers' provision of two or more versions of the same test to permit several assessments of the same skills or behaviors (as in a pre-post or longitudinal study with the same group of students). The use of alternate forms reduces concerns that students' scores may improve solely as a consequence of "learning the test" from repeated administration of the same items. To demonstrate that both forms of the test are essentially equivalent, a group of students takes both forms of the test (the time between administrations may vary). Alternate form reliability is demonstrated if the scores on the two forms are highly correlated. (See Reliability.)

Alternative assessment. Assessments administered using different approaches to assess a construct by using the same or similar content (portfolio review, written or oral responses to open-ended questions, journal review). Sometimes the term refers to an assessment designed to measure a construct among students with disabilities who are unable to take the standard assessment. (See Construct, Performance-based assessment.)

Analytic scoring. A type of rubric scoring that rates different dimensions rather than an overall evaluation of performance. A score is assigned individually to each dimension based on qualitative descriptions of performance on that dimension. The individual scores are sometimes combined for an overall score. For example, a mathematics assessment that uses analytic scoring may include items covering three concepts, with a score given for the quality of mastery of each concept and these scores may be summed for an the overall score. This sum of scores contrasts with a holistic rubric that would use a single rubric to rate the overall quality of mastery of these mathematics concepts. (See Rubric.)

Assessment. A method of gathering information about and/or quantifying an individual's performance in a particular topic area or on a specific task.

Authentic assessment. A method of assessing student performance that relies on tasks or probes similar to those encountered in daily activities or life. (See Performance-based assessment.)

Basal. With items ordered by difficulty in an adaptive test, the lowest item at which there is strong confidence (usually greater than 95 percent likelihood) that all previous items would be answered correctly by the individual if the test were administered. Adaptive tests provide rules regarding the number of consecutive items that must be answered correctly in order to establish a basal. Raw scores are the sum of (1) all items before and including the basal item, and (2) the number of correct items up to the ceiling item. (See Adaptive test, Ceiling, Floor effect)

Battery. Either a group of tests designed and marketed together as a comprehensive tool to assess a range of achievement or performance indicators, such as all academic skills, or a set of assessments that provide a measure of a specific area, such as different assessments used to identify a reading problem. Sometimes a developer collects and standardizes data on the same sample for all the tests in a battery so that results across tests can be compared (for example, the Woodcock-Johnson-III [WJ-III] Tests of Achievement and Tests of Cognitive Ability make up a psycho-educational battery). At other times, battery is used to refer to any group of tests administered (though not standardized) together to provide information on individual ability or achievement (for example, educational psychologists may administer a group of different reading and language tests to a student to evaluate reading difficulties).

Benchmark. Detailed criterion-referenced thresholds used to monitor a student's progress toward meeting performance standards, usually established according to age or grade levels based on state standards or other requirements. For example, correctly reading 20 basic sight words might be an end of kindergarten benchmark in reading. A benchmark may also be established as reaching a certain proficiency level on a standardized test. In the latter case, item response theory (IRT) may be used to determine an appropriate score as benchmark. (See Performance standards.)

Bias analysis. Characteristics of an assessment that unfairly favor one or more groups of children on the basis of factors such as gender, urban/rural residence, socioeconomic status, race/ethnicity, culture, or language. In a statistical context, bias reflects a systematic error in scores that compromises the generalizability of the results to a broader population. One common statistical procedure for examining bias of assessment items is differential item functioning. (See Differential item functioning [DIF], Generalizability.)

Ceiling. With items ordered by difficulty in an adaptive test, the highest item at which there is strong confidence (usually greater than 95 percent likelihood) that all subsequent items would be answered incorrectly by the individual if those items were administered. Adaptive tests provide rules regarding the number of items (usually consecutive items) that must be answered incorrectly in order to establish a ceiling. Raw scores are the sum of (1) all items before and including the basal item and (2) the number of correct items up to the ceiling item. (See Adaptive, Basal.)

Ceiling effect. Inability of an instrument to detect a difference in performance at the highest level, either among the highest performers or among all performers because an assessment was too easy. A ceiling effect can limit the ability to detect differences in groups across time points.

Classical test theory. Theory used for most test development in the last century. Classical test theory states that the observed score is equal to the true score (the latent ability) plus error. The theory assumes that the error is distributed normally and uniformly among students, has an

expected value of 0, and is not correlated with any variables. The item discrimination and difficulty in classical test theory (that is, point-biserial correlations and p values) are dependent upon the distribution of abilities in the sample so that large representative samples are needed to establish item properties. In classical test theory, these parameters are fixed and cannot be separated for an individual score. (See Latent trait).

Composite score. A score derived from several scores on interrelated tests based on use of a specific formula. For example, the WJ-III derives composite scores in reading comprehension from the performance of students on separate WJ-III tests.

Concurrent validity. Demonstration of the association (usually measured as a correlation) between a score on a given measure and performance on another assessment of the same or similar construct obtained at approximately the same time. (See Construct, Convergent, Criterion-related validity, Divergent validity)

Confidence interval. An interval surrounding a statistic (such as a mean, percentile, or correlation), within which the true statistic is believed to lie with a specified level of confidence (for example, with a 95 percent confidence level).

Construct. The trait to be assessed (for example, mathematics ability, empathy, social competence, reading achievement, or intelligence). The construct is a concept or characteristic of a student, teacher, or classroom that an assessment is supposed to measure.

Construct irrelevance. The extent to which scores on an assessment are influenced by factors irrelevant to the construct. Such factors distort the intended meaning of scores, which would have otherwise been accurate had the measurement been conducted under ideal circumstances (for example, if a student has no interest in and little prior knowledge about sports and all the passages on a reading test are thematically tied to sports, the student would not do as well on the reading test as students with an interest in and prior knowledge of sports).

Construct validity. Estimate of the degree to which an assessment measures the theoretical construct it claims to measure and to which inferences based on the assessment are relevant to the construct. Different sources of evidence support estimates of construct validity including evidence of a positive relationship with other measures of that construct or a similar construct (convergent validity) and expected weak or negative relationships with other constructs (divergent or discriminant validity). Evidence of construct validity also includes criterion-related validity evidence that demonstrates a relationship between the assessment score and performance on a task that is an independent measure of some skill related to the construct, such as receiving a passing grade in the subject area being tested. (See Convergent validity, Criterion-related validity, Divergent validity.)

Constructed response. An assessment requiring students to construct their own response rather than select an answer from a list of choices. Some in the assessment field believe that a constructed response is a truer test of student ability; others believe that multiple-choice tests are more reliable because answers are scored more objectively with less potential of scoring error or bias associated with assessor subjectivity.

Contamination. Systematic variance that distorts measurement of the construct. For example, administering a reading assessment that includes a science article to some students before they take the science assessment could lead to higher scores in science for those students than for students in a class that took the reading test after the science test.

Content domain. A defined body of knowledge and/or set of tasks, activities, or personal characteristics (including those with practical interrelationships). For example, the Compendium includes measures that assess a variety of content domains (as depicted in Table A.1).

Content sampling. Selection of areas or dimensions of content that are expected to represent the content domain. (See Content domain.)

Content standard. An outcome statement specifying what every child should know and be able to do in a particular content area. Content standards define not only what is expected of students but also what is expected of schools. In contrast, a performance standard says how well a child should be able to demonstrate knowledge and skills and gauges the degree to which a child has met a content standard. (See Performance standards.)

Content validity. An indicator providing information about whether a measure includes items relevant to and representative of the construct it is supposed to assess. No statistics are associated with content validity. Instead, the indicator is based on the professional judgment of experts who review the items to verify that the measure represents the domain that the developer intended and that the items provide variety and a range of difficulty. (See Construct.)

Convergent validity. A type of construct validity providing evidence of a positive relationship with other measures of that or a similar construct. This may be evaluated by looking at bivariate correlations between measures or the evidence may include the use of factor analysis that demonstrates that items in similar measures load on the same construct (while items in other measures load on different constructs demonstrating divergent validity). (See Construct, Divergent validity.)

Correlation. The degree to which two sets of scores or other data vary together, ranging from -1.0 (a perfect negative relationship) to 1.0 (a perfect positive relationship), with 0 indicating no association.

Criterion. A definition of acceptable performance levels or specific behaviors or pre-established level of mastery. Criteria may be used to develop scoring rubrics for assessments or to determine levels of performance for items on a measure such as "stands on one foot for 10 seconds". (See Rubric.)

Criterion-referenced assessment. An assessment measuring a student's performance against a specific criterion instead of against the performance of other students or a norming group. It is possible that no students or all students reach the pre-established expected level of mastery of a topic or skill. (See Norming sample.)

Criterion-related validity. The extent to which scores on an assessment are statistically related to a criterion (such as promotion to the next grade) or to scores on some other measure (preferably a well-respected or established measure) of the same objectives or criteria. It includes

both concurrent validity (taken at same time) and predictive validity (the criterion measured in the future).

Cronbach's coefficient alpha. An estimate of internal consistency reliability that is, how well groups of items on an assessment "hang together" or measure a particular trait or characteristic because of common factors among them. The greater the covariance among items, the higher the reliability is (and thus the higher the value of Cronbach's coefficient alpha). Values of the alpha can range from -1.0 to 1.0 with greater values indicating stronger internal consistency. The Cronbach's coefficient alpha is an extension of Kuder-Richardson Formula 20 (KR-20), a measure of internal consistency that is used when the items are dichotomous (right/wrong). (See KR-20 Kuder-Richardson Formula 20).

Cutoff score. A score that determines an acceptable minimum level of performance to pass an assessment or meet a standard.

Derived score. A score resulting from the transformation of the original raw score on an assessment. Examples of derived scores include standard scores, normed scores, or IRT scale scores. (See Raw score.)

Diagnostic assessment. A test that identifies a student's individual areas of weakness or strength and, where possible, uses the pattern of responses to identify the source of any weaknesses.

Differential item functioning (DIF). A statistical property of a test item. DIF arises when different groups of test takers with the same overall ability on the trait being tested demonstrate differences in how they perform on an item according to their particular group membership (for example, male versus female, white versus Hispanic). DIF is based on the principle that when different groups of test takers have roughly the same skill level (for example, mathematics knowledge), they should perform similarly on individual assessment items regardless of group membership. Typically, comparisons are based on gender, race, and language, but others are possible. Items demonstrating significant DIF often undergo review by content experts and may be removed from the measure if their inclusion unfairly favors one group over another.

Discriminant analysis. Statistical analysis to determine if a measure discriminates between students, teachers, or classrooms with different expected levels on the measured trait (for example, a student with a disability in reading should score lower on a reading assessment than students without a disability).

Divergent validity (sometimes referred to as **Discriminant validity**). Evidence of a weaker or absent relationship between two measures intended to represent different constructs (for example, a lack of a significant relationship between the student's score on a science assessment and ratings of the student's social interaction). Divergent validity may also be demonstrated by a strong negative relationship between two constructs (for example, a measure of problem behaviors would be negatively associated with a measure of cooperation. (See Construct, Convergent validity.)

Dynamic assessment. A method of assessment that involves examining not only whether a student perform a particular task but also what level of support is needed for the student to

succeed. This interactive approach, usually used in clinical contexts, assesses how the student responds to intervention.

Equivalent forms. Alternate forms of an assessment lacking the statistical similarity of parallel forms, but that compensate for dissimilarities in raw score statistics through conversions to derived scores or form-specific norm tables. (See Alternate forms, Parallel forms.)

Expert rater. A person who meets or exceeds the criteria required for competence in conducting an assessment (for example, a classroom observation of mathematics instruction quality). Inter-rater reliability may be determined as agreement among pairs/groups of raters or in relation to the expert rater. (See Inter-rater reliability, Rater.)

Factor analysis. Statistical analysis that examines the pattern of relationships among items in related groups, using correlations or a covariance matrix. Factor analysis may be exploratory (looking at how items group together in the data) or confirmatory (examining whether the relationships among items are consistent with a predetermined factor structure). (See Correlation.)

Fairness. Examination of how fairly an assessment measures a construct across diverse groups. Students from different demographic backgrounds should have an equal opportunity to demonstrate mastery on an assessment. Fairness in testing and assessment is indicated by equitable treatment, similarity of predictive validity for different groups of students/teachers, and absence of item bias after careful scrutiny for possible bias when subgroup differences are observed. Differential item functioning (DIF) analysis is one way to evaluate fairness across groups. (See Bias analysis, Differential item functioning (DIF).)

Floor effect. Inability of a test to assess some students because the test is too difficult; that is, there are not enough easy items to discriminate among students. If a student cannot answer the easiest items on an assessment, the assessment cannot measure the student's abilities in the construct.

Fluency. Observed ability to respond easily and quickly when, for example, naming letters or words or reading sentences. Fluency also refers to procedural fluency as in ease of implementing procedures such as addition, subtraction, or problem solving. Measures of fluency are often timed assessments.

Generalizability Theory (G-theory). Measure of the reliability and accuracy of results under different conditions. By recognizing the several sources of error in assessments, G-theory estimates how reliably a score on an assessment represents a student's level on a given trait. G-theory has also been applied to examining validity coefficients across several studies.

Grade equivalent score. Scores based on the median raw score of students of a given grade in the norming sample. The grade equivalent score is the median score for students in a particular grade that corresponds to the students' raw score, based on student performance in the norming sample on the same test. A grade equivalent score can be misleading in that an assessment's content may not include content across all grade levels for which a grade-equivalent score is estimated. (See Age equivalent, Norming sample, Raw score.)

Growth scale value. Transformations of students' IRT scores that create positive values that maintain a continuous interval scale. The values are more easily interpreted than a raw or standard score when examining longitudinal change. (See Raw score, W-score.)

Internal consistency reliability. A measure of the reliability of a score derived from the relationship among items of a single instrument and their ability to measure the same construct. Internal consistency reliability is presented as the correlation between groups of items or among all items. For example, split-half reliability refers to the correlation between the odd- and even-numbered items in an assessment. Another measure of internal consistency reliability is based on the correlations among all individual assessment items such as Cronbach's alpha or Kuder-Richardson Formula 20 (KR-20). (See Cronbach's coefficient alpha, KR-20, Split-half reliability.)

Inter-rater reliability. Extent to which different raters or observers obtain the same information; it can include agreement on scoring of items, administrative procedures, or observation of a given behavior. It is usually reported as either the correlation between the scores or ratings obtained by two observers or the percentage of items on which the two agree. Developers may also use an intra-class correlation (ICC) to compare the variance between raters to the total variance in the ratings. In research, inter-rater reliability is often a certification criterion for assessors/observers that must be met at the conclusion of training and during in-field data collection. (See Correlation, Inter-rater reliability, Rater, Reliability.)

Intraclass correlation (ICC). When used as a measure of inter-rater reliability, the ratio of the variance due to the independent variable (trait) divided by the explained variance plus the residual variance due to rater differences and measurement error. The ICC is sensitive to the sample of students assessed, such that samples with more restricted variance will have lower reliability estimates than those with greater variance.

IRT. See Item response theory model.

Item. A statement, question, exercise, or task on an assessment or measure.

Item difficulty index. The proportion of the test group that provides an incorrect response on a given item. In using classical test theory with achievement tests, an item difficulty range between 20 and 80 is preferable because items with values above or below that range do not perform well for the majority of the population. An average item difficulty index score between 30 and 60 is optimal for discriminating performance across students.

Item discrimination index. Ability of an item to discriminate between students of different ability or proficiency levels in relation to the difficulty of the item. In classical test theory, the item discrimination index is the difference between the proportion of the upper group that responded correctly to an item and the proportion of the lower group that responded correctly to the same item. The index may reach a maximum value of 100 for an item with an index of difficulty of 50, that is, when 100 percent of the upper group and none of the lower group answer the item correctly. For items of less than or greater than 50 difficulty, the index of discrimination has a maximum value of less than 100. (See Item, Item difficulty index.)

Item response theory (IRT) model. A method of producing scale scores based on a set of principles of measurement that result in estimates not biased by the sample distribution of ability. IRT uses information from all of the items and all of the students to estimate the item difficulties and the person abilities on the same scale. IRT models use the responses of all students to all of the questions to estimate the item difficulties and the person abilities on the same scale. The student's score on the measure is the estimate of the item difficulty at which the student has a 50 percent probability of answering the item correctly.

Kuder-Richardson Formula 20 (KR-20). A derivation of Cronbach's coefficient alpha that is used when items are dichotomous (right/wrong). KR-20 is often used as an indicator of internal consistency. Values can range from 0 to 1.0 with higher values indicating stronger internal consistency. The length of the assessment, variance in scores and the difficulty of the test can influence the KR-20. (See Cronbach's coefficient alpha.)

Latent trait (latent ability). A construct that cannot be directly observed, for example, intelligence, mathematics ability, empathy. Assessment use observations of indicators of the level of a latent trait to measure that trait.

Likert scale. A type of rating scale that assesses varying levels of student performance, student or teacher behavior, or classroom quality. It allows respondents to indicate their agreement or endorsement of a questionnaire statement. For example, a Likert scale may be used to assess student perceptions of safety in school. Response categories may range in number (for example, a four-point scale may range from strongly agree, agree, or disagree to strongly disagree).

Linear interpolation. Method for providing standard scores for shorter intervals between assessment periods. For example, a test may have standard scores based on an annual administration of the test, and researchers may wish to interpolate the scores for a six-month period. Linear interpolation assumes that growth in the content domain or skill is linear. (See Content domain, Standard scores.)

Measurement bias. See Bias analysis.

Measurement error. Error in measuring a variable that results from one or many factors including characteristics of an individual taking and administering a test, scoring accuracy, testing environment, and test administration. Error may be random or systematic and may affect the measure's reliability and validity.

Meta-analysis. A statistical method that synthesizes the results from several independent studies of comparable phenomena to estimate the strength of the relationship between variables.

Normal curve equivalent (NCE). A standard score with a mean of 50 and a standard deviation of 21.06. The NCE ranges from 1 to 99 and is a conversion of percentile rank into an equal-interval scale, making the NCE more suitable than percentiles for comparisons relative to the sample or to a normative sample.

Normative. Pertaining to information from a norming sample on which descriptive statistics or score interpretations are based (for example, percentiles, or expectancy). (See Norming sample).

Norming sample. The group of students whose scores on an assessment are used to establish the standardized scoring system, or norms for the assessment. Norming samples are selected to be representative of the population of interest, usually of the population of students in the United States based on recent census data. (See Representativeness of norming sample).

Norms. The distribution of expected scores obtained from the norming sample (see above) that describes performance on a particular assessment relative to the average of those in the sample. Norms typically serve to represent a larger population.

Objective. Pertaining to scores obtained in a way that minimizes bias or error due to different observers or scorers. Also used to refer to statements describing the performance or traits to be assessed.

Observer. A person who conducts a classroom observation. Also referred to as an assessor or rater.

Parallel forms. Alternate forms that have equal raw score means, equal standard deviations, equal error structures, and equal correlations with other measures for any given population. (See Alternate form.)

Percentile rank. Indicates a score's relative ranking in units 0 to 100 to other scores in a sample, usually a nationally representative norming sample. Interpretation is based on the percentage of students in the norming sample that performed in a similar way. A student whose score is at the 65th percentile has scored higher than 65 percent of the students in the norming sample. However, caution should be taken in comparing percentiles to each other because the raw score difference between percentiles will vary depending on the percentiles' location and the distribution of scores. In other words, percentiles are not on an equal interval scale. Normal curve equivalents convert percentile ranks into equal interval scores for ease of comparison of performance over time and across assessments. (See Normed scores, Normal curve equivalent, Norming sample).

Performance-based assessment. Evaluation of a student's comprehensive knowledge by asking them to complete "real life practical and intellectual challenges," in a setting that mirrors daily life as closely as possible. Performance-based assessments require students to construct a response or complete a task instead of selecting an answer from a list. Formats include exhibitions, investigations, demonstrations, written and/or oral responses to questions, journals, and portfolios. Performance is often gauged using a predetermined rubric, and may be assessed over time. Also called alternate, alternative, or authentic assessment.

Performance standards. Definitions stating what a student must do to reach various levels (for example, exceptional or above average) of performance on a specific task or test. The standards aim to gauge how well a student is doing on pre-established content standards. (See Benchmark).

Population. The universe of relevant cases from which a sample is drawn and to which the sample results may be generalized.

Practice effects. Influence of taking a test on subsequent performance on the same or a similar test, usually inflating performance due to familiarity with the specific questions. These effects are most evident when the time between administrations is short and the test is the same. Practice effects are one reason publishers provide alternate forms of assessments.

Predictive validity. Indicator of a type of criterion-related validity that demonstrates how accurately scores from a measure can predict scores on another measure or criteria assessed or gathered in the future. Researchers and assessment developers determine whether the measure is correlated with later functioning. If the correlation between the two measures obtained across the time interval is high, evidence of predictive validity is established. If, for example, a measure of vocabulary in kindergarten is highly correlated with an assessment of reading ability in second grade, the vocabulary assessment could be said to have evidence of predictive validity. In some cases, researchers use other activities or events as the criterion, rather than another assessment. For example, researchers might show a positive correlation between kindergarten vocabulary and second grade language arts report card grades as evidence of predictive validity. In general, the younger the student being assessed, the poorer the predictive validity of the assessment. (See Criterion-related validity).

Psychometrics. The study of psychological or educational measurement in areas such as knowledge, aptitude, attitude, skill, and the quality of the care and education environment. Psychometric properties document evidence that indicates how reliable and valid a measure is based on the purposes for which it was designed and used.

Quotient. A conversion of a raw score that describe the performance of an individual relative to the average performance of a sample. Most often used in the context of intelligence testing and referred to as intelligence quotient or IQ. (See Normed score, Standardized score).

Random sample. Set of individuals or cases selected according to a randomized process. A random sample is usually statistically representative of the population from which it was drawn.

Rasch model (Rasch-based scores). A latent trait model, also considered a one-parameter item response theory (IRT) model. Rasch models assume single trait is measured and equal item discrimination. Rasch models estimate the student scores in relation to the difficulty of the items. Rasch-based scores are equal interval with both the student scores and the item difficulties estimated on the same scale. These scores are expressed in logits that have positive and negative values, and so are often transformed to have positive values (See W Scores, Growth Scale Value).

Rater. A person who evaluates student performance on a test or conducts a classroom observation. Also referred to as assessor or observer.

Rating scale. Values assigned to varying levels of student performance, student or teacher behavior, or classroom quality. These could be in the form of numbers, descriptions, or the absence or presence of a characteristic. For example, the rating scale for items on a test generally have two levels, correct or incorrect. A Likert scale is a type of rating scale that allows

respondents to indicate their agreement or endorsement of a questionnaire statement across multiple categories of agreement. Rating scales may also rate frequency or quality and can have many categories. For example, a student might rate enjoyment of reading from 1 "not at all" to 10 "most enjoyable activity that I know".

Raw score. Total number of correct or endorsed items on an assessment of student performance, student or teacher behavior, or classroom quality. Raw scores are used to create other scores such as standardized scores, percentile ranks, age- or grade-equivalents, and item response theory (IRT) scores.

Reliability. The extent to which scores obtained from an assessment or group of assessments are accurate and consistent over one or more possible sources of error, including time, raters, items, environment, and sample groups of a population. Indicators of reliability assess how dependable a measure is for the purpose it is used. Reliable measures are stable over time and include items that measure the same thing in different ways. For tools that require standardized observation (for example, classroom quality observations or ratings of student's behavior), the scores obtained by two different, well-trained observers must be similar to be considered reliable. Statistical measures of reliability are typically reported as coefficients, which range from -1.0 to 1.0, with a greater value reflecting greater reliability. Many researchers and assessment developers require that assessment and screening tools have reliability values of 0.7 or higher. Typical indicators of reliability include alternate form, internal consistency, inter-rater, and test-retest. An unreliable assessment cannot be valid. (See Internal consistency, Inter-rater reliability, Test-retest reliability).

Representativeness of norming sample. Degree to which a sample used to standardize a measure represents the participants in the study sample, in the nation or some other population of interest. Most developers include information about the norming sample in their manuals.

Restriction of range or variability. Reduction in the observed score variance of a sample, compared to the variance of an entire population, because of constraints on the process of sampling, for example by sampling only students who are exactly five years of age at the beginning of kindergarten and have gone to a prekindergarten program compared with sampling all kindergartners. The correlation between the performances of these students on two different assessments would be constrained by the limited variance in abilities of these students. Variance may be restricted because the subsample of the population is very homogenous, or if a measure has ceiling or floor problems; that is, the sample of items does not match the ability level of the students and the variance in the results is restricted. When looking for relationships between a sample with restricted variance and one that has greater variability, the strength of the correlations will be constrained by the limited variance.

Rubric. A guide or scale used during an assessment to gauge student performance, teacher behavior, or classroom quality according to the description of each level on the scale. The rubric sets specific criteria required to achieve each level and is often used with subjective assessments (for example, essay writing). A rubric can be holistic and rate the overall quality or it can be analytic with individual rubrics for each dimension.

Sample. A selection of a specified number students or teachers from a larger set of people called the population.

Scale scores. Any of several scores converted from original raw scores on an assessment. Typically the scale scores range from 1 to 999, with equal intervals between scores, making them useful to for making comparisons. Scale scores are designed to compare the performance of an individual to the full distribution of a sample or population, or to him/herself across time. Item response theory (IRT) scale scores, norm-referenced and standard scores are examples of scale scores.

Sensitivity. A characteristic of screening tests that indicates the proportion of students at risk for some disability or difficulty who are correctly identified by the test. Lower values (<.80) indicate a problem with under-referral (under-identification) of the disability/difficulty for further assessment.

Specificity. A characteristic of screening tests that indicates the proportion of students who are correctly identified as not at risk for a disability or difficulty. Lower values (<.80) indicate a problem with over-referrals (incorrectly identified) for additional assessment.

Split-half reliability. A form of internal consistency reliability, obtained by splitting the items on an assessment in half and obtaining two independent scores. The correlation between these two scores, usually adjusted using the Spearman-Brown formula (derived from classical test theory), provides an estimate of the reliability of the entire assessment.

Standards-based assessment. An assessment in which the criteria for item selection are derived directly from content and/or performance standards.

Standard error of measurement (SEM). The standard deviation of an individual's observed scores from repeated administrations of an assessment under identical conditions. The SEM is typically estimated from group data (rather than from repeated measures from a single person) and can be interpreted as the precision or reliability of scores on the assessment. Every assessment has a different SEM for a given sample of students. If a test had perfect reliability (i.e., no measurement error), the SEM would equal zero.

Standard score. A derivation of the raw score (raw scores are usually the sum of the correct or observed items) obtained on a measure in relation to the distribution of the norming sample scores. Standard scores are expressed in standard units. Thus, the difference in performance between standard scores of 85 and 90 is the same as the difference between that of 55 and 60. Standard scores have specified values for the mean and standard deviation, for example, z-scores have a mean of 0 and a standard deviation of 1.0, while quotients used for many standardized assessments have a mean of 100 and a standard deviation of 15 (quotients). (See T-scores, Quotients, and Normal curve equivalents (NCEs).

Standardization. Use of a uniform or standard set of procedures for administering and scoring an assessment. This may include both establishing scoring norms based on performance of the norming sample, and maintaining a consistent test environment during its administration. (See Norming sample, Representativeness of the norming sample.)

Stanine score. Provide information on students' performance relative to students in the norming sample, much like percentile ranks. Stanines divide the normal curve into nine intervals (range 1 to 9), with the lowest scores falling into the first stanine, the highest scores falling into

the ninth stanine, and the fifth stanine straddling the midpoint of the distribution. Stanine scores have a mean of 5 and a standard deviation of 2. Except for the two extreme stanines (the first and the ninth), the range of each stanine is one-half of a standard deviation unit. A disadvantage of stanine scores is that they magnify small differences between raw scores that fall on either side of a point separating adjacent stanines.

Statistical significance. The finding that empirical data are inconsistent with a null hypothesis, usually that no difference exists between groups, at some specified probability level.

Stratified random sample. A set of random samples from several different subsets, or strata, of the population, grouped by a set of common characteristics (for example, gender, race/ethnicity, and age). Stratified random samples are statistically representative of the population from which they are drawn, and ensure that a pre-specified proportion of the sample comes from each stratum included in the population (See Random sample.)

Subscale (also called **Subtest).** A set of items within a larger assessment that measure a particular aspect of the trait being measured. Scores for a particular subtest are often the unit of analysis in educational research. For example, separate tests of reading comprehension and fluency may be subscales within an overall reading assessment. Subscales may be specified based on theoretical grounds (grouping items based on their content) or empirical evidence (factor analysis of items in a longer scale may reveal meaningful subscales).

Test. A procedure used to observe and evaluate a student's performance or behavior in a specific area or task, which is then summarized by a score. Tests can be norm- or criterion-referenced. (See Assessment).

Test or assessment development. Process through which a test, observation, or other type of measure is planned, constructed, evaluated, and modified. Evaluation and design elements include content, format, administration, scoring, item properties, scaling, and technical quality for its intended purpose.

Test information function (TIF). A measure of reliability available from Item Response Theory (IRT) models. The TIF provides estimates of the amount of information provided by a test at different levels of proficiency and thus provides more accurate estimations (than a reliability coefficient) of how well a given trait is measured at different levels of the trait. The TIF allows examination of the precision across the range of the ability distribution. If the test is supposed to measure average achievement for a sample, the greatest precision should be near the mean of the scale with acceptable values within two standard deviations of the mean. Typically, longer tests and well-designed adaptive tests provide greater precision and have stronger TIF.

Test-retest reliability. The stability of test results over time. Evidence of test-retest reliability involves testing the same group of individuals at least twice, with a relatively short interval between assessments, usually no longer than a few days or weeks apart. The reliability coefficient is then obtained by correlating both sets of scores. The higher the test-retest reliability, the more stable the assessment tool is considered to be. Longer periods between administrations of the same assessment will reduce the reliability, partly because the individual's situation (for example, skill) can be expected to change. Some also consider a measure to be test-

retest reliability when a student is tested on different forms of the same test. (See Alternate-form reliability.)

Theta. An estimate of the student's level of the trait or construct, sometimes referred to as the "person ability score." Psychometricians use Item Response Theory (IRT) models to estimate the theta based on the individual's responses and characteristics of the items on an assessment (such as item difficulty and discrimination). (See Item Response Theory [IRT])

T-score. A standard score with a mean of 50 and a standard deviation of 10.

Validity. The degree to which an assessment accurately measures what it is designed to measure. Validity is often measured in comparison to other instruments established to measure the same or similar behavior/traits. Types of validity include content, construct, and predictive. An assessment cannot be valid if it is not reliable.

W-scores (**W-ability scores**). A linear transformation of the person ability score (the theta in Item Response Theory [IRT] models originally expressed in logits). The W-score is an example of a growth scale value and is the appropriate score to use in looking at change over time. The W-score is designed to help with interpretability. When the W-score matches the item difficulty, it means that the individual has a 50 percent probability of correctly answering the question. With a 10-point increase in the W-score, the probability that the person could correctly answer that question increases to 75 percent, and a 20-point increase means that the probability is 90 percent. (See Growth scale value, Item Response Theory, Theta.)

SOURCES

Abrami, Philip C., Paul Cholmsky, and Robert Gordon. *Statistical Analysis for the Social Sciences: An Interactive Approach*. Boston: Allyn & Bacon, 2000.

Association for Supervision and Curriculum Development. "Lexicon of Learning." Available at [http://www.ascd.org/Publications/Lexicon_of_Learning/Lexicon_of_Learning.aspx]. 2008.

Baker, Frank. *The Basics of Item Response Theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, 2001.

Castle Rock Research. "Support Resource for Implementation of the Computer Adaptive Assessment Initiative—Alberta Computer Adaptive Assessment Glossary." Available at [http://projects.cbe.ab.ca/sss/ilscommunity/archive08/t/i/CAAGuideRevisions.pdf]. 2006.

Council of Chief State School Officers. "Early Childhood Glossary Terms." Available at [http://www.ccsso.org/projects/SCASS/projects/early_childhood_education_assessment_consortium/publications_and_products/2840.cfm]. 2008.

Cozby, Paul C. "Methods in Behavioral Research, Ninth Edition." Available at [http://highered.mcgraw-hill.com/sites/0073531812/student_view0/chapter9/glossary.html]. 2007.

Harris, Theodore L., and Richard E. Hodges (eds.). *The Literacy Dictionary: The Vocabulary of Reading and Writing*. Newark, DE: International Reading Association, 1995.

Howitt, Dennis, and Duncan Cramer. "Introduction to Research Methods in Psychology, Second Edition, Companion Website." Available at [http://wps.pearsoned.co.uk/ema_uk_he_howitt_resmethpsy_2/77/19812/5071921.cw/index.html]. 2007.

Instructional Assessment Resources. "Glossary." Available at [http://www.utexas.edu/academic/diia/assessment/iar/glossary.php]. 2007.

LDLD Project. "Glossary of Assessment Terms." Available at [http://www.ldldproject.net/glossary.html]. 2005.

Litwin, Mark. *How to Assess and Interpret Survey Psychometrics*. Thousand Oaks, CA: SAGE Publications, 2003.

McAfee, Oralie, Deborah J. Leong, and Elena Bodrova. "Glossary." In *Basics of Assessment: A Primer for Early Childhood Educators*, edited by Bry Pollack. Washington, DC: National Association for the Education of Young Children, 2004.

National Center for Research on Evaluation, Standards, and Student Testing. "CRESST Glossary." Available at [http://www.cse.ucla.edu/products/glossary.html]. n.d.

New Horizons for Learning. "Assessment Terminology: Glossary of Useful Terms." Available at [http://www.newhorizons.org/strategies/assess/terminology.htm]. 2002.

Ohio Department of Mental Health. "Research Glossary." Available at [http://mentalhealth.ohio.gov/what-we-do/promote/research-and-evaluation/learning-lab/research-glossary.shtml]. 1999.

Ohio Resource Center for Mathematics, Science, and Reading. "Basic Data & Value-Added Glossary of Terms." Available at [http://www.ohiorc.org/orc_documents/orc/value-added/documents/GLOSSARY.doc]. 2008.

Online Evaluation Resource Library. "OERL Definition of Terms." Available at [http://oerl.sri.com/definitions.html]. n.d.

Pearson Assessments. "BASI Growth Scale Value (GSV) Scales and College Report Manual Addendum." Bloomington, MN: Pearson Assessments. Available at [http://www.pearsonassessments.com/pdf/basi_addendum.pdf]. n.d.

Pearson Assessments. "Glossary." Available at [http://www.ed.pearsonassessments.com/glossary.php]. 2008.

Pearson Assessments. "Qualification Levels and Requirements." Available at [http://pearsonassessments.com/catalog/qualification.htm]. 2007.

Quality Assurance in Language for Specific Purposes. "Glossary." Available at [http://www.google.com/url?sa=X&start=0&oi=define&ei=8uzcSJjTLImIeargoJ0E&sig2=gw1lgRjNnTGEMCb-aK85Iw&q=http://www.qalspell.ttu.ee/Glossary.doc&usg=AFQjCNHTi4hkxTFIlK8lUw9rpnQ6V8bCEw]. 2003.

Society for Industrial and Organizational Psychology. "Glossary of Terms." Available at [http://www.siop.org/_principles/pages66to72.pdf]. n.d.

Statistics.com. "Statistical Glossary." Available at [http://www.statistics.com/resources/glossary/]. 2007.

System for Adult Basic Education Support. "Glossary of Useful Assessment Terms." Available at [http://www.sabes.org/assessment/glossary.htm]. 2008.

Wisconsin Education Association Council. "Assessment and Testing Terms." Available at [http://www.weac.org/professional_resources/Testing/performance_assessment/test_glossary.aspx]. 2006.

Wright, B.D., and J. M. Linacre. "Glossary of Rasch Measurement Terminology." Available at [http://www.rasch.org/rmt/rmt152e.htm]. 2001.

APPENDIX F

CROSS-WALK OF OFFICIAL NCEE OR REL STUDY NAMES, ABBREVIATED NAMES, AND WEB ADDRESSES

TABLE F.1

CROSS-WALK OF OFFICIAL NCEE OR REL STUDY NAMES, ABBREVIATED NAMES, AND STUDY WEB ADDRESSES

Official NCEE or REL Study Name[a]	**NCEE or REL Study Abbreviated Name**	**Web Address**
A Multisite Cluster Randomized Trial of the Effects of Compass Learning Odyssey Math on the Math Achievement of Selected Grade 4 Students in the Mid-Atlantic Region	Odyssey Math® (REL-Mid-Atlantic)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=26
A Study of Classroom Literacy Interventions and Outcomes in Even Start Classrooms	Even Start Classroom Literacy	http://ies.ed.gov/ncee/pubs/20084028/
Accelerating language development in kindergarten through Kindergarten PAVEd for Success	Accelerating language development (REL-Southeast)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=67
An Evaluation of Teachers Trained Through Different Routes to Certification	Different Routes to Certification	http://ies.ed.gov/ncee/pubs/20094043/pdf/20094043.pdf
An Evaluation of the Impact of Mandatory Random Student Drug Testing	Mandatory Random Student Drug Testing	http://ies.ed.gov/ncee/projects/evaluation/drugtesting.asp
An Investigation of the Impact of the 6 + 1 Trait® Writing Model on Student Achievement	6+1 Trait Writing Model (REL-Northwest)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=52
Assessing the Impact of Collaborative Strategic Reading (CSR) on Reading Comprehension	Collaborative Strategic Reading (REL-Southwest)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=78
Closing the Reading Gap	Closing the Reading Gap	http://ies.ed.gov/ncee/projects/evaluation/literacy_readinggap.asp
Differential Effects of English language learner training and materials—On Our Way to English (OWE) and Responsive Instruction for Success (RISE)	English Language Learner Training and Materials (REL-Central)	http://ies.ed.gov/ncee/projects/evaluation/literacy_readinggap.asp

Table F.1 *(continued)*

Official NCEE or REL Study Name[a]	NCEE or REL Study Abbreviated Name	Web Address
Effectiveness of a Small Group Mathematics Intervention for Struggling First Graders	Small Group Mathematics (REL-Southwest)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=80
Effectiveness of Reading and Mathematics Software Products	Reading and Mathematics Software Products	http://ies.ed.gov/ncee/pubs/20094041/pdf/20094041.pdf
Effects of the Lessons in Character English Language Arts Character Education Program on Behavior and Academic Outcomes	Lessons in Character Education (REL-West)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=91
Efficacy of Frequent Formative Assessment for Improving Instructional Practice and Student Performance, Given Variations in Training to Use Assessment Results	Formative Assessment (REL-Midwest)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=36
Evaluating the Impact of the Program for Infant/Toddler Care	Program for Infant/Toddler Care (REL-West)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=90
Evaluation of Early Elementary Math Curricula	Math Curricula	http://www.mathematica-mpr.com/publications/pdfs/education/mathcurricula_firstgradefind09.pdf
Evaluation of Principles-Based Professional Development to Improve Reading Comprehension for English Language Learners	Principles-Based Professional Development (REL-Pacific)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=61
Evaluation of Reading Comprehension Programs	Reading Comprehension	http://ies.ed.gov/ncee/pubs/20094032/pdf/20094032.pdf
Evaluation of the DC Opportunity Scholarship Program	DC Opportunity Scholarship	http://ies.ed.gov/ncee/pubs/20074009/

Table F.1 *(continued)*

Official NCEE or REL Study Name[a]	NCEE or REL Study Abbreviated Name	Web Address
Evaluation of the Quality Teaching for English Learners Program	Quality Teaching for English Learners (REL-West)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=88
Impact of the Thinking Reader Software Program on Grade 6 Reading Comprehension, Vocabulary, Strategies, and Motivation	Thinking Reader (REL-Northeast & Islands)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=46
Impact Evaluation of a School-Based Violence Prevention Program	School-Based Violence Prevention	http://ies.ed.gov/ncee/projects/evaluation/violence.asp
Impact Evaluation of the U.S. Department of Education's Student Mentoring Program	Student Mentoring Program	http://ies.ed.gov/ncee/projects/evaluation/mentoring.asp
Impact of the Understanding Science Professional Development Model on Science Achievement of English Language Learner Students	Science Professional Development (REL-West)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=87
Impacts of a Problem-Based Instruction Approach to Economics on High School Students)	Problem-Based Economics (REL-West)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=89
Impacts of Comprehensive Teacher Induction	Comprehensive Teacher Induction	http://ies.ed.gov/ncee/pubs/20094034/
Improving Adolescent Literacy Across the Curriculum in High Schools (Content Literacy Continuum, CLC)	Adolescent Literacy Across the Curriculum (REL-Midwest)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=34
National Evaluation of Early Reading First	Early Reading First	http://ies.ed.gov/ncee/pubs/20074007/index.asp
Project CRISS Reading Program and Grade 9 Reading Achievement in Rural High Schools	Project CRISS Reading Program (REL-Northwest)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=53

Table F.1 *(continued)*

Official NCEE or REL Study Name[a]	NCEE or REL Study Abbreviated Name	Web Address
Project ELLA (English Language/Literacy Acquisition)	Project ELLA	http://ies.ed.gov/ncee/projects/evaluation/ell_projectella.asp
Reading First Impact Study	Reading First	http://ies.ed.gov/ncee/pdf/20094038.pdf
The Effect of Connected Mathematics 2 on Math Achievement in Grade 6 in the Mid-Atlantic Region	Connected Mathematics Program 2 (REL-Mid-Atlantic)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=25
The Effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI)	Alabama Math, Science, and Technology Initiative (REL-Southeast)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=69
The Effects of Classroom Assessment for Student Learning (CASL) on Student Achievement	Classroom Assessment for Student Learning (REL-Central)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=18
The Effects of Hybrid Algebra I on Teacher Practices, Classroom Quality, and Adolescent Learning	Hybrid Algebra I (REL-Appalachia)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=8
The Effects of Opening the World of Learning (OWL) on the Early Literacy Skills of At-Risk Urban Preschool Students	Opening the World of Learning (REL-Appalachia)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=9
The Effects of Success in Sight as a School Improvement Intervention	Effects of Success in Sight (REL-Central)	http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=20
The Enhanced Reading Opportunities Study	Enhanced Reading Opportunities	http://ies.ed.gov/ncee/pubs/20084015/index.asp
The Evaluation of Enhanced Academic Instruction in After-School Programs	Enhanced Academic Instruction in After-School Programs	http://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=NCEE20084021

Table F.1 *(continued)*

Official NCEE or REL Study Name[a]	NCEE or REL Study Abbreviated Name	Web Address
The Impact of Professional Development Strategies on Teacher Practice and Student Achievement in Math	Professional Development Strategies in Math	http://ies.ed.gov/ncee/projects/evaluation/tq_mathematics.asp
The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement	Professional Development Interventions on Early Reading	http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=NCEE20084034

[a] The official NCEE or REL study name was obtained from the study description page on the web site for the National Center for Education Evaluation and Regional Assistance (NCEE) or Regional Educational Laboratory Program (REL) within the Institute for Education Sciences, U.S. Department of Education. Each official study name is hyperlinked to its NCEE or REL web site or latest study report as of the time of the Compendium.

APPENDIX G

INDEX OF STUDENT ACHIEVEMENT/DEVELOPMENT MEASURES INCLUDED IN THE COMPENDIUM, BY CATEGORY

TABLE G.1

PAGE NUMBER FOR STUDENT ACHIEVEMENT/DEVELOPMENT MEASURES INCLUDED IN THE COMPENDIUM, BY CATEGORY

Category	Measure	Format[a]	Page
Comprehensive Cognitive and Achievement Tests	Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III)	Profile	B.21
	Kaufman Test of Educational Achievement, Comprehensive Form, Second Edition (KTEA-II)	Profile	B.93
	Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP) and Achievement Level Tests (ALT)	Profile	B.119
	Stanford Achievement Test Series, Tenth Edition (Stanford-10)	Profile	B.197
	TerraNova 3	Profile	B.209
	Woodcock Johnson-III Normative Update (WJ III NU)	Profile	B.253
Literacy, Reading	6+1 Trait Writing Scoring Guide (Rubrics)	Profile	B.3
	AIMSweb Oral Reading Fluency	Profile	B.7
	Dynamic Indication of Basic Early Literacy Skills (DIBELS), 6th Edition	Profile	B.33
	Gates-MacGinitie Reading Test, Fourth Edition (GMRT-4)	Profile	B.63
	Group Reading Assessment and Diagnostic Evaluation (GRADE)	Profile	B.69
	Indicadores Dinámicos del Éxito en la Lectura (IDEL), 7th Edition	Profile	B.87
	Metacognitive Awareness of Reading Strategies Inventory (MARSI)	Profile	B.109
	Phonological Awareness Literacy Screening (PALS) PreK, PALS-K, PALS 1-3	Profile	B.137
	Science Reading Comprehension Assessment	Profile	B.169
	Social Science Reading Comprehension Assessment	Profile	B.185
	Stanford Diagnostic Reading Test, 4th Edition (SDRT)	Profile	B.203
	Test of Preschool Early Literacy (TOPEL; formerly PreCTOPP)	Profile	B.231

Table G.1 (*continued)*

Category	Measure	Format[a]	Page
	Test of Silent Contextual Reading Fluency (TOSCRF)	Profile	B.235
	Test of Silent Word Reading Fluency (TOSWRF)	Profile	B.241
	Test of Word Reading Efficiency (TOWRE)	Profile	B.247
	Woodcock Reading Mastery Test-Revised/Normative Update (WRMT-R/NU)	Profile	B.261
Vocabulary, Communication	Expressive One-Word Picture Vocabulary Test—Third Edition (EOWPVT)	Profile	B.51
	Expressive Vocabulary Test, Second Edition (EVT-2)	Profile	B.57
	Lexical diversity	Table	B.266
	MacArthur-Bates Communicative Development Inventories (CDI)	Profile	B.101
	Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4)	Profile	B.131
	Preschool Individual Growth and Developmental Indicators (IGDI)	Profile	B.151
	Preschool Language Scale, 4th Edition (PLS-4)	Profile	B.157
	Test of Language Development - Primary, Fourth Edition (TOLD-4)	Profile	B.225
Language Proficiency	IDEA Oral Language Proficiency Test (IPT I-Oral English)	Profile	B.75
	IDEA Oral Language Proficiency Test, 3rd Edition (IPT I-Oral Spanish)	Profile	B.81
	PRELAS 2000	Profile	B.145
Mathematics	Algebra End-of-Course Assessment	Profile	B.13
	Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K) Mathematics Assessment	Profile	B.43
	Test of Early Mathematics Ability, Third Edition (TEMA-3)	Profile	B.217
Science	Assessing Teacher Learning about Science Teaching (ATLAST) Test of Force and Motion	Profile	B.17
Social Studies	Student Performance Assessment Tasks (UCLA/CRESST)	Table	B.269

Table G.1 (*continued)*

Category	Measure	Format[a]	Page
	Test of Economic Literacy, Third Edition (TEL-3)	Profile	B.221
Approaches toward Learning, Motivation	Motivation for Reading Questionnaire (MRQ)	Profile	B.113
	Patterns of Adaptive Learning Scales (PALS)	Profile	B.125
	The Research Assessment Package for Schools – Student Self Report (RAPS)	Profile	B.163
	Self- and Task-Perception Questionnaire	Profile	B.173
	Student questionnaire of economic interest and attitudes	Table	B.266
	Student Time-on-Task and Engagement with Print (STEP)	Table	B.267
Social-Emotional Well-Being	Social Competence and Behavior Evaluation, Preschool Edition (SCBE)	Profile	B.179
	Social Skills Rating System (SSRS)	Profile	B.189
	Student questionnaire of behaviors and violence	Table	B.268
Other/Multidomain	Character traits and behavior questionnaire	Table	B.269
	Student questionnaire of reading behavior and attitudes	Table	B.270
	Student questionnaire of substance use	Table	B.271
	Student questionnaire on behavior and school	Table	B.271

Note: The Compendium includes these student achievement/development measures because of their use as an outcome in a recent NCEE or REL study evaluating an educational intervention using randomized controlled trials or quasi-experimental design.

[a] Format refers to how the Compendium presents the available information. A profile includes an overview and narrative. The table format summarizes important information about measures recently developed, generally those for studies with less technical information available to support completion of a full profile at the time of Compendium development.

TABLE G.2

PAGE NUMBER FOR TEACHER KNOWLEDGE MEASURES INCLUDED IN THE COMPENDIUM, BY CATEGORY

Category	Measure	Format[a]	Page
Reading Knowledge (content and/or pedagogical)	Reading Content and Practices Survey (RCPS)	Table	C.20
	Teacher impact questionnaire of ELL instructional pedagogy	Table	C.21
Mathematics Knowledge (content and/or pedagogical)	Pedagogical Content Knowledge Assessment (PCK)	Profile	C.11
	Teacher Knowledge Inventory (TKI)	Table	C.22
Science Knowledge (content and/or pedagogical)	Assessing Teacher Learning about Science Teaching (ATLAST) Test of Force and Motion	Profile	B.17
Social Studies Knowledge	Test of Economic Literacy, Third Edition (TEL-3)	Profile	B.221
Pedagogical Knowledge	Test of assessment knowledge	Table	C.22
Multidomain Pedagogical Content Knowledge	Diagnostic Classroom Observation Tool (DCO; formerly VCOT)	Profile	C.5
	Reformed Teaching Observation Protocol (RTOP)	Profile	C.15

Note: The Compendium includes these teacher knowledge measures because of their use as an outcome in a recent NCEE or REL study evaluating an educational intervention using randomized controlled trials or quasi-experimental designs.

[a]Format refers to how the Compendium presents the available information. A profile includes an overview and narrative. The table format summarizes important information about measures recently developed, generally those for studies with less technical information available to support completion of a full profile at the time of Compendium development.

TABLE G.3

PAGE NUMBER FOR CLASSROOM PRACTICES AND SETTING MEASURES INCLUDED IN THE COMPENDIUM, BY CATEGORY

Category	Measure	Format[a]	Page
Comprehensive Classroom Practices	Authentic Instructional Practices Classroom Observation Form	Profile	D.3
	CIERA classroom observation scheme for classroom literacy instruction	Profile	D.13
	Classroom Characteristics (CC) form	Table	D.82
	Diagnostic Classroom Observation Tool (DCO, formerly VCOT)	Profile	C.5
	Early Childhood Environment Rating Scale-Revised Edition (ECERS-R)	Profile	D.21
	Early Reading Professional Development (PD) Classroom Observation	Profile	D.33
	Infant/Toddler Environment Rating Scale, Revised Edition (ITERS-R)	Profile	D.39
	Reformed Teaching Observation Protocol (RTOP)	Profile	C.15
	School Observation Measure (SOM)	Profile	D.59
	Sheltered Instruction Observation Protocol (SIOP)	Profile	D.65
	Teacher questionnaire of attitudes and behaviors	Table	D.82
	Teacher questionnaire of classroom quality and instructional practices	Table	D.83
	Teacher questionnaire of educational practices	Table	D.83
Reading Practices	Classroom observations of instructional quality	Table	D.85
	Classroom observation of literacy teaching practices	Table	D.85
	Early Language & Literacy Classroom Observation (ELLCO) Pre-K and K-3 Tools	Profile	D.27
	Expository Reading Comprehension Classroom Observation (ERCCO)	Table	D.86
	Instructional Practice in Reading Inventory (IPRI)	Table	D.87
	Lexical diversity	Table	D.88
	Literacy Observation Tools (LOT; E-LOT, LOT, and A-LOT)	Profile	D.45
	Observation Measure of Language and Literacy Instruction (OMLIT)	Profile	D.51
	Teacher Behavior Rating Scale (TBRS)	Profile	D.71
	Teacher Interaction and Language Rating Scale	Profile	D.77

Table G.3 (*continued)*

Category	Measure	Format[a]	Page
	Teacher questionnaire on reading instructional practices	Table	D.88
Mathematics Practices	Algebra I Quality Assessment (AQA)	Table	D.89
	Algebra I teacher questionnaire	Table	D.90
	Classroom observation of math practices	Table	D.91
	Observation of Math Instruction (OMI) form	Table	D.91
School Engagement or Climate	Student and parent questionnaires of school climate	Table	D.92
	Student questionnaire of behaviors and violence	Table	D.93
	Teacher questionnaire of school climate	Table	D.93
	Teacher questionnaire on safety and victimization	Table	D.94
Other/Multidomain	Caregiver Interaction Scale (CIS)	Profile	D.9
	Student questionnaire of economic interests and attitudes	Table	D.94
	Teacher questionnaire of instructional practices	Table	D.95
	Teacher questionnaire of instructional practices and self-efficacy	Table	D.95
	Teacher questionnaire of practices and economic attitudes	Table	D.96

Note: The compendium includes these classroom practices and setting measures because of their use as an outcome in an NCEE or REL study evaluating an educational intervention using randomized controlled trials or quasi-experimental designs.

[a] Format refers to how the Compendium presents the available information. A profile includes an overview and narrative. The table format summarizes important information about measures recently developed, generally those for studies with less technical information available to support completion of a full profile at the time of Compendium development.