ED 353 302 TM 019 366

AUTHOR Mislevy, Robert J.
TITLE Linking Educational Assessments: Concepts, Issues, Methods, and Prospects.
INSTITUTION Educational Testing Service, Princeton, NJ. Policy Information Center.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Dec 92
CONTRACT R117G10027
NOTE 97p.; Foreword by Robert L. Linn.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Academic Achievement; Competence; *Educational Assessment; Elementary Secondary Education; Equated Scores; Evaluation Utilization; *Measurement Techniques; Projective Measures; Statistical Analysis; Student Evaluation
IDENTIFIERS Calibration; *Linking Educational Assessments; Statistical Moderation

ABSTRACT

This paper describes the basic concepts of linking educational assessments. Although it discusses statistical machinery for interpreting evidence about students' achievements, this paper's principal message is that linking tests is not just a technical problem. Technical questions can not be asked or answered until questions about the nature of learning and achievement and about the purposes and consequences of assessment ~re addressed. Some fundamental ideas about educational assessment and test theory (chains of inferences and roles of judgment) are considered. The following approaches to linking assessments are described: (1) equating in physical measurement and educational assessment; (2) calibration in physical measurement and educational assessment; (3) projection in physical measurement and educational assessment; and (4) statistical moderation and social moderation. Implications for a system of monitoring progress toward educational standards are discussed. The following obtainable goals are highlighted: (1) comparing levels of performance across clusters directly in terms of common indicators of performance on a selected sample of consensually defined tasks administered under standard conditions; (2) estimating levels of performance of groups or individuals within clusters at the levels of accuracy demanded by purposes within clusters; (3) comparing levels of performance across clusters in terms of performance ratings on customized assessments in terms of a consensually defined, more abstract description of developing competence; and (4) making projections about how students from one cluster might have performed on the assessment of another cluster. Seven tables, 14 figures, and 41 references are included. (RLC)

POLICY INFORMATION CENTER

Educational
Testing Service

Linking Educational Assessments

Concepts, Issues, Methods, and Prospects

Linking Educational Assessments:
Concepts, Isues, Methods, and Prospects

By

Robert J. Mislevy
Educational Testing Service

With a Foreword by

Robert L. Linn
Center for Research on Evaluation,
Standards, and Student Testing
University of Colorado at Boulder

December 1992

ETS
Policy Information Center
Princeton, NJ 08541

Table of Contents

Page

List of Tables and Figures

Page

Tables

Figures

Preface

As this is written, a fairly wide consensus has developed around the proposition that we need to establish national education standards and a system of assessments to measure whether the standards are being achieved. Rejecting a single national examination, the call is for a voluntary system of assessments, all geared to national standards, in which states, groups of states, or other groups design their own assessments. According to the report of the National Council on Education Standards and Testing, the key features of assessments "would be alignment with high national standards and the capacity to produce uscful, comparable results."

We focus here on the word "comparable." Irrespective of which assessment is given, there is a desire to be able to translate the results so that students across the country can be compared on their achievement of the national standards. Indeed, this has become a key to developing consensus. To those who fear that national standards and an assessment system will lead to a national curriculum, the response is that the standards are just "frameworks" within which curriculum can be determined locally, and assessments can reflect this curriculum. Techniques would be used, such as "calibration," to make the results comparable. At a meeting of the National Education Goals Panel, a governor said something to the effect that we can have as many tests as we want; all we have to do is calibrate them. How far can local curriculum stray from the national standards and differ from one locality to another, and how different can assessments be in how and what they test and still enable comparable scores to be constructed?

The ETS Policy Information Center commissioned Robert J. Mislevy to give guidance to the policy and education community on this question. While his report establishes the range of freedom, which may not be as wide as many have assumed or hoped, it also tries to give guidance on how comparability *can* be achieved. It is critical, I believe, to know this in *advance* and design the system in a way that comparability is possible. There are no neat technical tricks by which the results of just any old assessments can be compared.

We asked Mislevy to write in as nontechnical a manner as possible so that a person with a need to know could understand. However, it is not a simple matter. The conclusion, on page 72, will give the less motivated reader the bottom line on what Mislevy thinks possible, along with a summary of approaches to linking tests.

We are grateful to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) for joining in the funding of this work, and we are grateful for the willingness of Robert Linn to write the foreword; his own work has already contributed much to the understanding of both limitations and possibilities.

Paul E. Barton, Director
Policy Information Center

Acknowledgments

This monograph evolved from an internal memorandum on linking assessments by Martha Stocking and me. The work received support from the Policy Information Center and the Statistical and Psychometric Research Division of Educational Testing Service and the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Educational Research and Development Program, cooperative agreement number R117G10027 and CFDA catalog number 84.117G, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

I am grateful to many colleagues for discussions on the topics addressed in this paper and comments on earlier drafts, among them Nancy Allen, Paul Barton, Al Beaton, Charles Davis, John Frederiksen, Gene Johnson, Charlie Lewis, Bob Linn, Skip Livingston, John Mazzeo, Sam Messick, Carol Myford, Mimi Reed, Karen Sheingold, Martha Stocking, Kentaro Yamamoto, Howard Wainer, Michael Zieky, and Rebecca Zwick. The views I express do not necessarily represent theirs. Joanne Pfleiderer and Richard J. Coley provided editorial assistance, Jim Wert designed the cover, and Carla Cooper was responsible for desktop publishing services.

Robert J. Mislevy

Foreword

As is attested to by the abundance of concordance tables that have been developed to provide conversions of composite scores of the American College Testing (ACT) Program Assessment to the scale of the Scholastic Aptitude Test (SAT) or vice versa, the interest in linking results from different assessments is not new. Some of the advocated changes in the nature and uses of educational assessments during the past few years, however, have greatly expanded interest in issues of linking results of different assessments. The January 1992 report of the National Council on Education Standards and Testing* (NCEST) clearly illustrates the need for the type of careful attention to issues of linking that is found in Bob Mislevy's report.

The NCEST report recommended the development of national "content standards" and "student performance standards" that would define the skills and understandings that need to be taught and the levels of competence that students need to achieve. The Council also concluded that a system of assessments was needed to make the standards meaningful. Two features of the system of assessments envisioned by the Council are particularly relevant to the issues addressed by Mislevy. First, the Council recommended that the national system of assessments involve multiple assessments rather than a single test. As we stated in the report, states or groups of states would be expected "to adopt assessments linked to national standards. States can design the assessments or they may acquire them" (NCEST, 1992, p. 30). California might develop its own assessments, Arizona might contract with a test publisher to develop its own assessments, and a group of New England states might join forces to develop a common set of assessments for use in those states. But it is hoped that they would all be linked to the national standards.

Second, as is implicit in the desired linking to national standards, the Council concluded that "it is essential that different assessments produce comparable results in the attainment of the standards" (NCEST, 1992, p. 30). That is, it is expected that the performance of students who respond to one set of assessment tasks in Florida and that of students who respond to a completely different set of tasks in Illinois can nonetheless be compared in terms of common national standards. It is expected that a "pass" or "high pass" will have common meaning and value in terms of the national standards despite the use of different assessments in different locales.

Another example of the expanded desire for linking comes from states that want to express state assessment results in terms of the scales used by the National Assessment of Educational Progress (NAEP). The state assessment may differ from NAEP in format and in its content specifications, but there is still a desire to estimate the percentage of students in the state who would score above a given level on NAEP based on their performance on the state assessment.

*The National Council on Education Standards and Testing. (1992). *Raising standards for American education*. Washington, DC: Author.

The basis for linking assessments that might be developed according to the NCEST plan or for a state to link its results to NAEP is quite different from that undergirding the equating of alternate forms of a test developed using common test specifications by a single test publisher. Yet, some of the discussion of the desired linkings blurs the distinctions with the use of a wide variety of terminology, some of which is undefined in a technical sense and some of which is used in ways quite inconsistent with well-established technical meanings. Confusion has resulted from the use of terms such as equating, calibration, benchmarking, anchoring, moderation, verification, and prediction with little regard for the implied technical requirements or for specific types of comparisons that are justified.

Bob Mislevy's paper provides much needed clarifications of the terminology. More importantly, it provides a lucid explication of the concepts of equating, calibration, projection, and statistical moderation along with concrete illustrations of the important distinguishing characteristics of these concepts and the associated statistical methods. By using physical examples such as various measures of temperature, Mislevy clearly illustrates that the issues involved in justifying various types of comparisons are not unique to the peculiarities of assessments of student knowledge, skills, and understandings. Rather, the requirements for the simpler and more rigorous forms of linking result "not from the statistical procedures used to map the correspondence, but from the way the assessments are constructed" (page 73).

Mislevy's paper should prove to be of great value in clarifying the discussion of linking issues. Although not everyone will like the fact that some desired types of correspondence for substantially different assessments are simply impossible, the paper points the direction to "attaining less ambitious, but more realistic, goals" (page 73) for linking different assessments. The report provides a much needed foundation that should help the field achieve greater specificity about attainable linking goals and lead to the use of techniques suitable to support more realistic interpretations of results from different assessments.

Robert Linn
Center for Research on Evaluation,
Standards, and Student Testing
University of Colorado at Boulder

Introduction

Can We Measure Progress Toward National Goals if Different Students Take Different Tests?

The problem of how to link the results of different tests is a century old—and as urgent as today's headlines. Charles Spearman first addressed aspects of this problem in his 1904 paper, "The proof and measurement of association between two things."

This monograph was inspired by the President's and governors' recent joint statement on *National Goals for Education:*

> *The time has come for the first time in the United States history to establish clear national performance goals, goals that will make us internationally competitive.*
>
> *[N]ational educational goals will be meaningless unless progress toward meeting them is measured accurately and adequately, and reported to the American people.*
>
> *National goals for education,*
> U.S. Department of Education,
> 1990

> *The academic performance of elementary and secondary students will increase significantly in every quartile, and the distribution of minority students will more closely reflect the student population as a whole.*
>
> *The percentage of students who demonstrate the ability to reason, solve problems, apply knowledge, and write and communicate effectively will increase substantially.*
>
> Two objectives for Goal #3,
> student achievement and
> citizenship, in *National*
> *goals for education, pp. 5*

The specific language of these objectives seems to demand a common yardstick for measuring student achievement across the nation. But recognizing that different schools and communities may well want to assess different competencies in different ways,

> *...the National Educational Goals Panel, the National Council on Education Standards and Testing, and the New Standards Project ... all discourage the use of a*

> *single national exam...They advocate that ... districts or states wishing to participate in the new system should form "clusters" agreeing to develop common exams linked to the national standards. Performances on the assessments would then be "calibrated" against the national standards to yield comparable data for different exams used by different clusters. ... Proponents of a new national system of exams ... say that the "cluster" model of calibrating the results of different tests to common standards will avoid the problem of dictating to states and local districts what should be taught.*
>
> *ASCD Update,* 1991, 33(8), pp. 1-6

Can technical "calibration procedures" use results from different assessments to gauge common standards? Professor Andrew Porter is skeptical:

This paper describes the basic concepts of linking educational assessments.

> *If this practice of separate assessments continues, can the results be somehow equated so that results on one can also be stated in terms of results on the other? There are those who place great faith in the ability of statisticians to equate tests, but that faith is largely unjustified. Equating can be done only when tests measure the same thing.*
>
> Andrew Porter, 1991, pp. 35

Technical questions cannot be asked—much less answered—until questions about the nature of learning and achievement and about the purposes and consequences of assessment are dealt with.

This paper describes the basic concepts of linking educational assessments. Although it discusses statistical machinery for interpreting evidence about students' achievements, the principal message is that linking tests is not just a technical problem. Technical questions cannot be asked—much less answered—until questions about the nature of learning and achievement and about the purposes and consequences of assessment are dealt with. We begin by discussing some fundamental ideas about educational assessment and test theory. We then describe and illustrate approaches to linking assessments. Finally, we consider implications for a system of monitoring progress toward educational standards.

Educational Assessments

I use the term "educational assessment" to mean systematic ways of gathering and summarizing evidence about student competencies. It includes, but is not limited to, familiar standardized tests in which large numbers of students answer the same kinds of prespecified questions under the same conditions, such as the Scholastic Aptitude Test (SAT), the National Assessment of Educational Progress (NAEP), and the written part of driver's license exams. I also include such activities as doctoral dissertations and the portfolios students create for the College Board's Advanced Placement (AP) Studio Art Examination. These latter kinds of assessments require concentrated effort over an extended period of time on challenges determined to a large degree by the student. Summarizing evidence on these assessments requires judgment.

Assessments provide data such as written essays, correct and incorrect marks on an answer sheet, and students' explanations of the rationales for their problem solutions. This *data* becomes *evidence*, however, only with respect to conjectures about students' competence[1]—perhaps concerning individual students, the group as a whole or particular subgroups, or even predictions about future performance or the outcomes of instruction. Our purpose for an assessment and our conception of the nature of competence drive the form an assessment takes. An assessment that produces solid evidence at reasonable costs for one mission can prove unreliable, exorbitant, or simply irrelevant for another.

An assessment that produces solid evidence at reasonable costs for one mission can prove unreliable, exorbitant, or simply irrelevant for another.

Suppose that Assessment X provides evidence about instructional or policy questions involving student competencies. In particular, we might want to know whether individuals can perform at particular levels of competence—i.e., whether they are achieving specified educational standards. Suppose that with appropriate statistical tools, we could provide answers, each properly qualified by an indication of the strength of evidence.

Let's also suppose that someone wants to link Assessment Y to Assessment X. This notion stems from a desire to address the same questions posed in terms of Assessment X when we observe students' performances on Assessment Y rather than on X. The degree to which linking can succeed, and the nature of the machinery required to carry it out, depend on the matchups between the purposes for which the assessments were constructed and the aspects of competence they were designed to reveal.

The degree to which linking can succeed, and the nature of the machinery required to carry it out, depend on the matchups between the purposes for which the assessments were constructed and the aspects of competence they were designed to reveal.

[1] Schum (1987, p. 16) emphasizes the distinction between data and evidence in the context of intelligence analysis. His book provides a wide variety of examples that offer insights into inferential tasks in educational assessment.

Purposes of Assessment

A wide variety of purposes can motivate educational assessment, and we make no effort to survey them here. Table 1 gives the interested reader a feel for some of the ways assessment purposes might be classified. What's important is that different kinds and amounts of evidence must be gathered to suit different purposes. Certain distinctions among purposes prove especially pertinent to our discussion about linking:

- Will important decisions be based on the results; that is, is it a "high-stakes" assessment? A quiz to help students decide what to work on during a particular day is a low-stakes assessment—a poor choice is easily remedied. An assessment to determine whether students should graduate from eighth grade is high stakes at the individual level. One designed to decide how to distribute state funds among schools is high stakes at the school level. Assessments that support important decisions must provide commensurately dependable evidence, just as a criminal conviction demands proof "beyond a reasonable doubt."

- Do inferences concern a student's comparative standing in a group of students—that is, are they "norm-referenced?" Or do they gauge the student's competencies in terms of particular levels of skills or performance—that is, are they "criterion referenced?" A norm-referenced test assembles tasks to focus evidence for questions such as "Is Eiji more or less skilled than Sung-Ho?" A criterion-referenced test in the same subject area might include similar tasks, but they would be selected to focus evidence for questions such as "What levels of skills has Eiji attained?"

- Do inferences concern the competencies of individual students, as with medical certification examinations, or the distributions of competencies in groups of students, as with NAEP? When the focus is on the individual, enough evidence must be gathered on each student to support inferences about him or her specifically. On the other hand, a bit of information about each student in a sample—too little to say much about any of them as an individual—can suffice in the aggregate to monitor the level of performance in a school or a state.

Table 1[1]
Description of Assessment Purposes

	Domain to Which Inferences Will Be Made				
	Curricular Domain				
Type of Inference Desired	**Before Instruction**	**During Instruction**	**After Instruction**	**Cognitive Domain**	**Future Performance in Criterion Setting**
Description of individual examinees' attainments	Placement	Diagnosis	Grading	Reporting to students and/or parents	Guidance and counseling
Mastery decision (individual examinee above or below cutoff)	Selection	Instructional guidance	Promotion	Certification Licensing	Selection Admission Licensing
Description of performance for a group or system	Preinstruction status for research or evaluation	Process and curriculum evaluation	Postinstruction status for research or evaluation Accountability reporting at the level of, e.g., schools or instructional programs	Construct measurement for research or evaluation Accountability reporting at state or national levels	Research and planning; e.g.: • determining cutoff levels for mastery decisions • evaluating the effectiveness of instructional programs

[1] Adapted from Millman & Greene's Table 8.1 (1989)

Conceptions of Competence

The role of psychological perspectives on competence in educational assessment can be illustrated by two contrasting quotations. The first reflects the behaviorist tradition in psychology:

> *The educational process consists of providing a series of environments that permit the student to learn new behaviors or modify or eliminate existing behaviors and to practice these behaviors to the point that he displays them at some reasonably satisfactory level of competence and regularity under appropriate circumstances. The statement of objectives becomes the description of behaviors that the student is expected to display with some regularity. The evaluation of the success of instruction and of the student's learning becomes a matter of placing the student in a sample of situations in which the different learned behaviors may appropriately occur and noting the frequency and accuracy with which they do occur.*
>
> D.R. Krathwohl & D.A.
> Payne, 1971, pp. 17-18

From the behaviorist perspective, the specifications for an assessment describe task contexts from the assessor's point of view and provide a system for classifying student responses.

From the behaviorist perspective, the specifications for an assessment describe task contexts from the assessor's point of view and provide a system for classifying student responses. Responses in some contexts are unambiguously right or wrong; in other contexts, the occurrence of certain type of behaviors, the classification of which may require expert judgment, are recorded.

The second quotation reflects what has come to be called a cognitive perspective:

> *Essential characteristics of proficient performance have been described in various domains and provide useful indices for assessment. We know that, at specific stages of learning, there exist different integrations of knowledge, different forms of skill, differences in access to knowledge, and differences in the efficiency of performance. These stages can define criteria for test design. We can now propose a set of candidate dimensions along which subject-matter competence can be assessed. As competence in a subject matter grows, evidence of a knowledge base that is increasingly* ***coherent, principled, useful, and goal-oriented*** *is displayed, and test items can be designed to capture such evidence. [emphasis original]*
>
> R. Glaser, 1991,
> pp. 26

From the cognitive perspective, the specifications for an assessment describe contexts that can evoke evidence about students' competence as conceived at a higher level of abstraction, and provide judgmental guidelines for mapping from observed behavior to this inferred proficiency. Table 2, the American Council on the Training of Foreign Languages (ACTFL) guidelines for assessing reading proficiency, illustrates the point. The ACTFL guidelines contrast Intermediate readers' performance using texts "about which the reader has personal interest or knowledge" with Advanced readers' comprehension of "texts which treat unfamiliar topics and situations." This distinction is fundamental to the underlying conception of developing language proficiency, but can a situation that is familiar to one student be obviously unfamiliar to others. The evidence conveyed by the same behavior in the same situation can differ radically for different students and alter what we infer about their capabilities from their behavior.

From the cognitive perspective, the specifications for an assess- ment describe contexts that can evoke evidence about students' competence as conceived at a higher level of abstraction, and provide judgmental guidelines for mapping from observed behavior to this inferred proficiency.

Assessments as Operational Definitions

Educators can agree unanimously that we need to help students "improve their math skills" but disagree vehemently about how to appraise these skills. Their conceptions of mathematical skills often diverge as they move from generalities to the classroom because they employ the language and concepts of differing perspectives on how mathematics is taught and learned as well as what topics and skills are important. Disparate assessments all provide evidence about students' competence—but each takes a particular point of view about what competence is and how it develops.

Disparate assessments all provide evidence about students' competence—but each takes a particular point of view about what competence is and how it develops.

Figure 1[2] is a hypothetical illustration of the levels that could lie between common, broad perceptions of educational goals and a variety of assessments. Higher in the scheme are generally-stated objectives, such as:

> *The percentage of students who demonstrate the ability to reason, solve problems, apply knowledge, and write and communicate effectively will increase substantially.*
>
> *National goals for education,*
> U.S. Department of Education,
> 1990

[2] Suggested by Payne's discussion of levels of specificity in educational outcomes (1992, pp. 64).

Table 2
ACTFL Proficiency Guidelines for Reading*

Novice-Low	Able occasionally to identify isolated words and/or major phrases when strongly supported by context.
Novice-Mid	Able to recognize the symbols of an alphabetic and/or syllabic writing system and/or a limited number of characters in a system that uses characters. The reader can identify an increasing number of highly contextualized words and/or phrases including cognates and borrowed words, where appropriate. Material understood rarely exceeds a single phrase at a time, and rereading may be required.
Novice-High	Has sufficient control of the writing system to interpret written language in areas of practical need. Where vocabulary has been learned, can read for instructional and directional purposes standardized messages, phrases, or expressions, such as some items on menus, schedules, timetables, maps, and signs. At times, but not on a consistent basis, the novice-high reader may be able to derive meaning from material at a slightly higher level where context and/or extralinguistic background knowledge are supportive.
Intermediate-Low	Able to understand main ideas and/or some facts from the simplest connected texts dealing with basic personal and social needs. Such texts are linguistically noncomplex and have a clear underlying internal structure, for example, chronological sequencing. They impart basic information about which the reader has to make only minimal suppositions or to which the reader brings personal interest and/or knowledge. Examples include messages with social purposes or information for the widest possible audience, such as public announcements and short, straightforward instructions for dealing with public life. Some misunderstandings will occur.
Intermediate-Mid	Able to read consistently with increased understanding simple connected texts dealing with a variety of basic and social needs. Such texts are still linguistically noncomplex and have a clear underlying internal structure. They impart basic information about which the reader has to make minimal suppositions and to which the reader brings personal information and/or knowledge. Examples may include short, straightforward descriptions of persons, places, and things written for a wide audience.
Intermediate-High	Able to read consistently with full understanding simple connected texts dealing with basic personal and social needs about which the reader has personal interest and/or knowledge. Can get some main ideas and details from texts at the next higher level featuring description and narration. Structural complexity may interfere with comprehension; for example, basic grammatical relations may be misinterpreted and temporal references may rely primarily on lexical items. Has some difficulty with cohesive factors in discourse, such as matching pronouns with referents. While texts do not differ significantly from those at the Advanced level, comprehension is less consistent. May have to read several times for understanding.
Advanced	Able to read somewhat longer prose of several paragraphs in length, particularly if presented with a clear underlying structure. The prose is predominantly in familiar sentence patterns. Reader gets the main ideas and facts and misses some details. Comprehension derives not only from situational and subject matter knowledge but from increasing control of the language. Texts at this level include descriptions and narrations such as simple short stories, news items, bibliographical information, social notices, personal correspondence, routinized business letters, and simple technical material written for the general reader.

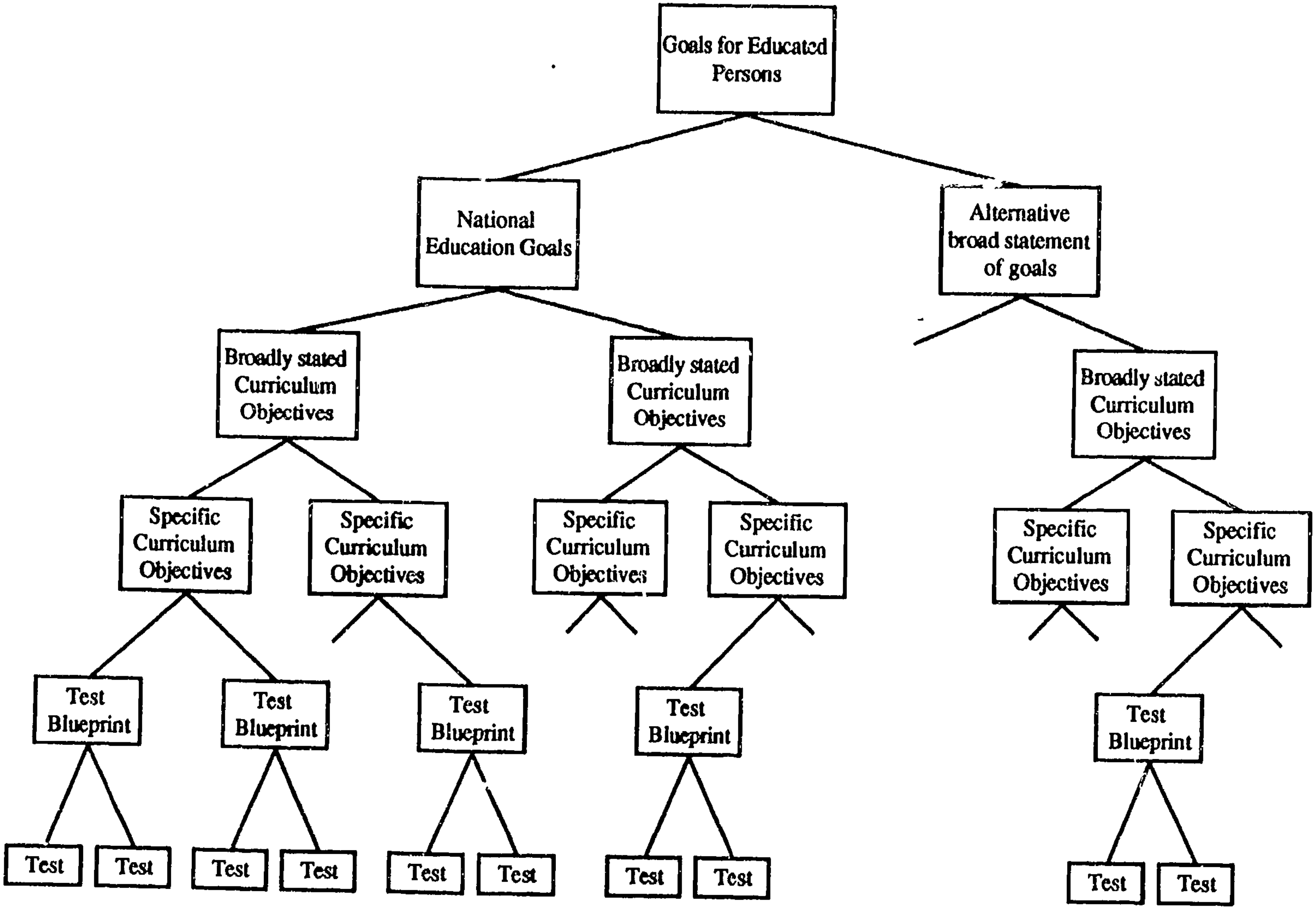
Table 2, continued
ACTFL Proficiency Guidelines for Reading

Advanced-Plus	Able to follow essential points of written discourse at the Superior level in areas of special interest or knowledge. Able to understand parts of texts which are conceptually abstract and linguistically complex, and/or texts which treat unfamiliar topics and situations, as well as some texts which involve aspects of target-language culture. Able to comprehend the facts to make appropriate inferences. An emerging awareness of the aesthetic properties of language and of its literary styles permits comprehension of a wider variety of texts, including literary. Misunderstandings may occur.
Superior	Able to read with almost complete comprehension and at normal speed expository prose on unfamiliar subjects and a variety of literary texts. Reading ability is not dependent on subject matter knowledge, although the reader is not expected to comprehend thoroughly texts which are highly dependent on the knowledge of the target culture. Reads easily for pleasure. Superior-level texts feature hypotheses, argumentation, and supported opinions and include grammatical patterns and vocabulary ordinarily encountered in academic/professional reading. At this level, due to the control of general vocabulary and structure, the reader is almost always able to match the meanings derived from extralinguistic knowledge with meanings derived from knowledge of the language, allowing for smooth and efficient reading of diverse texts. Occasional misunderstandings may still occur; for example, the reader may experience some difficulty with unusually complex structures and low-frequency idioms. At the superior level the reader can match strategies, top-down or bottom-up, which are most appropriate to the text. (Top-down strategies rely on real-world knowledge and prediction based on genre and organizational scheme of the text. Bottom-up strategies rely on actual linguistic knowledge.) Material at this level will include a variety of literary texts, editorials, correspondence, general reports, and technical material in professional fields. Rereading is rarely necessary, and misreading is rare.
Distinguished	Able to read fluently and accurately most styles and forms of the language pertinent to academic and professional needs. Able to relate inferences in the text to real-world knowledge and understand almost all sociolinguistic and cultural references by processing language from within the cultural framework. Able to understand the writer's use of nuance and subtlety. Can readily follow unpredictable turns of thought and author intent in such materials as sophisticated editorials, specialized journal articles, and literary texts such as novels, plays, poems, as well as in any subject matter area directed to the general reader.

* Based on the ACTFL proficiency guidelines, American Council on the Training of Foreign Languages (1989).

Figure 1

Hierarchy of Educational Outcomes


Goals for Educated Persons
National Education Goals
Alternative broad statement of goals
Broadly stated Curriculum Objectives
Broadly stated Curriculum Objectives
Broadly stated Curriculum Objectives
Specific Curriculum Objectives
Specific Curriculum Objectives
Specific Curriculum Objectives
Specific Curriculum Objectives
Specific Curriculum Objectives
Specific Curriculum Objectives
Test Blueprint
Test Blueprint
Test Blueprint
Test Blueprint
Test Blueprint
Test
Test
Test
Test
Test
Test
Test
Test
Test
Test

The National Council of Teachers of Mathematics' (NCTM) *Curriculum and Evaluation Standards for School Mathematics* offers steps along one possible path toward making such goals meaningful. The excerpt below and the examples in the NCTM *Standards* exemplify less abstract levels in Figure 1:

> *Standard 1: Mathematics as Problem Solving*
>
> In grades K-4, the study of mathematics should emphasize problem solving so that students can ...
>
> - use problem-solving approaches to investigate and understand mathematical content;
> - formulate problems from everyday and mathematical situations;
> - develop and apply strategies to solve a wide variety of problems;
> - verify and interpret results with respect to the original problem;
> - acquire confidence in using mathematics meaningfully.
>
> *Curriculum and Evaluation Standards for School Mathematics*, NCTM, pp. 23

The level of Figure 1 labeled "test blueprint" represents what a particular assessment should comprise: the kinds and numbers of tasks, the way it will be implemented, and the processes by which observations will be summarized and reported. This level of specificity constitutes an *operational definition* of competence. Quality-control statistician W. Edwards Deming describes how similar processes are routinely required in industry, law, and medicine:

> *Does pollution mean, for example, carbon monoxide in sufficient concentration to cause sickness in 3 breaths, or does one mean carbon monoxide in sufficient concentration to cause sickness when breathed continuously over a period of 5 days? In either case, how is the effect going to be recognized? By what procedure is the pres-*

> *ence of carbon monoxide to be detected? What is the diagnosis or criterion for poisoning? Men? Animals? If men, how will they be selected? How many? How many in the sample must satisfy the criteria for poisoning from carbon monoxide in order that we may declare the air to be unsafe for a few breaths, or for a steady diet?*
>
> *Operational definitions are necessary for economy and reliability. Without an operational definition, unemployment, pollution, safety of goods and of apparatus, effectiveness (as of a drug), side-effects, duration of dosage before side-effects become apparent (as examples), have no meaning unless defined in statistical terms. Without an operational definition, investigations on a problem will be costly and ineffective, almost certain to lead to endless bickering and controversy.*
>
> *An operational definition of pollution in terms of offensiveness to the nose would be an example. It is not an impossible definition (being close kin to statistical methods for maintaining constant quality and taste in foods and beverages), but unless it be statistically defined, it would be meaningless.*
>
> W. Edwards Deming,
> 1980, pp. 259

...an educational assessment often comprises multiple scores or ratings for each student, to provide a fuller picture of individual competencies. Because some linking methods are designed to align only one score with another, our discussion will focus on assessments that provide only a single score.

A study of pollution in cities might include several of these operational definitions, to provide a fuller picture of their environments. In the same way, an educational assessment often comprises multiple scores or ratings for each student, to provide a fuller picture of individual competencies. Because some linking methods are designed to align only one score with another, our discussion will focus on assessments that provide only a single score. In practice, this type of linking can proceed simultaneously among matching sets of scores from multifaceted assessments (for example, fractions scores with fractions scores, vocabulary scores with vocabulary scores, and so on).

Standardization

From any given set of specifications, an assessment can be implemented in countless ways. Differences, small or large, might exist in tasks, administration conditions, typography, identity and number of judges, and so on. *Standardizing* an aspect of an assessment means limiting the variations students encounter in that aspect, in an effort to eliminate classes of hypotheses about students' assessment results. Did Duanli score higher than Mark because she had more time, easier questions, or a more lenient grader? Standardizing

Did Duanli score higher than Mark because she had more time, easier questions, or a more lenient grader? Standardizing timing, task specifications, and rating criteria reduces the chance of this occurrence.

timing, task specifications, and rating criteria reduces the chance of this occurrence. Other potential explanations for the higher score exist, of course. The plausibility of each explanation can be strengthened or diminished with additional evidence.[3]

> Standardizing progressively more aspects of an assessment increases the precision of inferences about behavior in the assessment setting, but it can also reduce opportunities to observe students performing tasks more personally relevant to their own varieties of competence. Assessing developing competence when there is neither a single path toward "better" nor a fixed and final definition of "best" may require different kinds of evidence from different students (Lesh, Lamon, Behr, & Lester, 1992, pp. 407).

Note that "standardization" is not synonymous with "multiple-choice," although multiple-choice tests do standardize the range of responses students can make.

Note that "standardization" is not synonymous with "multiple-choice," although multiple-choice tests do standardize the range of responses students can make. The AP Studio Art portfolio—the antithesis of a multiple-choice test—is standardized in other respects. Among the requirements for each portfolio in the general section are four original works that meet size specifications; four slides focusing on color and design; and up to 20 slides, a film, or a videotape illustrating a concentration on a student-selected theme. These requirements ensure th~t evidence about certain multiple aspects of artistic development will be evoked, although the wide latitude of student choice virtually guarantees that different students will provide different forms of evidence.

Questions about which aspects of an assessment to standardize and to what degree arise under all purposes and modes of testing and all views of competence. The answers depend in part on the evidential value of the observations in view of the purposes of the assessment, the conception of competence, and the resource demands.

Test Theory

Test theory is the statistical machinery for reasoning from students' behavior to conjectures about their compe-

[3] A branch of test theory called generalizability (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) examines the impact of variations in aspects of assessment settings on test scores.

Test theory is the statistical machinery for reasoning from students' behavior to conjectures about their competence, as framed by a particular conception of competence.

tence, as framed by a particular conception of competence. In an application, a conception of competence is operationalized as a set of ways students might differ from one another. These are the variables in a student model, a simplified description of selected aspects of the infinite varieties of skills and knowledge that characterize real students. Depending on our purposes, we might isolate one or hundreds of facets. They might be expressed in terms of numbers, categories, or some mixture; they might be conceived of as persisting over long periods of time or apt to change at the next moment. They might concern tendencies in behavior, conceptions of phenomena, available strategies, or levels of development. We don't observe these variables directly. We observe only a sample of specific instances of students' behavior in limited circumstances—indirect evidence about their competence as conceived in the abstract.

Suppose we want to make a statement about Jasmine's proficiency, based on a variable in a model built around some key areas of competence. We can't observe her value on this variable directly, but perhaps we can make an observation that bears information about it: her answer to a multiple-choice question, say, or two sets of judges' ratings of her violin solo, or an essay outlining how to determine which paper towel is most absorbent. The observation can't tell us her value with certainty, because the same behavior could be produced by students with different underlying values. It is more likely to be produced by students at some levels than others, however. Nonsensically answering "¿Como está usted?" with "Me llamo Carlos," for example, is much more likely from a student classfied as a Low-Novice under the ACTFL guidelines than an Advanced student.

The key question from a statistician's point of view is "How probable is this particular observation, from each of the possible values in the competence model?" The answer—the so-called "likelihood function" induced by the response—embodies the information that the observation conveys about competence, in the way competence is being conceived. If the observation is equally likely from students at all values of the variables in the competence model, it carries no information. If it is likely at some values but not others, it sways our belief in those directions, with strength in proportion to how much more likely the observation is at those values.

For example, compare the results of observing two heads out of four coin tosses with observing 1,000 heads out of

2,000 tosses. Both experiments suggest that the most likely value of the probability of heads on any given toss is 1/2; that is, the coin is fair. The evidence from 2,000 tosses is much stronger, however. It would not be unusual to obtain two heads out of four tosses with a trick coin biased toward, say, a 3/4 probability of heads, while such a coin would hardly ever produce as few as 1,000 heads from 2,000 tosses.

The most familiar tools we have for linking assessments evolved under the paradigm of "mental measurement," introduced a century ago in an attempt to "measure intelligence."

The most familiar tools we have for linking assessments evolved under the paradigm of "mental measurement," introduced a century ago in an attempt to "measure intelligence." The measurement paradigm posits that important aspects of students' knowledge or skills can sometimes be represented by numbers that locate them along continua, much as their heights and weights measure some of their physical characteristics. From the behavioral point of view, the variable of interest might be the proportion of correct answers a student would give on every item in a large domain. From the cognitive point of view, the variable might be a level on a developmental scale such as the ACTFL reading guidelines. Standard test theory views observations as noisy manifestations of these inherently unobservable variables and attacks the problem of inference in the face of measurement error. We discuss classical test theory in the section about equating and discuss item response theory in the section about calibration. The statistical roots of those theories are grounded in physical measurement, so examples about temperature and weight illustrate the potential and the limitations of test theory to link assessments.

...the traits achievement tests purportedly measure, such as mathematical ability, reading level, or physics achievement, are not features of objective reality, but constructs of human design...

It is important to remember that the traits achievement tests purportedly measure, such as mathematical ability, reading level, or physics achievement, are not features of objective reality, but constructs of human design—invented to organize experience and solve problems, but shaped by the science and the society in which they evolved. Contemporary conceptions of learning do not describe developing competence in terms of increasing trait values, but in terms of alternative constructs: constructing and reconstructing mental structures that organize facts and skills ("schemas"); learning how to plan, monitor, and, when necessary, switch problem-solving strategies ("metacognitive skills"); and practicing procedures to the point that they no longer demand high levels of attention ("automaticity"). Test scores tell us something about what students know and can do, but any assessment setting stimulates a unique constellation of knowledge, skill, strategies, and motivation within each examinee, and different settings stimulate different unique constellations. The mental measurement paradigm is useful only to the extent that the patterns of behavior it captures guide instruction and policy among options framed in our current conceptions of how competence develops.[4]

The mental measurement paradigm is useful only to the extent that the patterns of behavior it captures guide instruction and policy among options framed in our current conceptions of how competence develops.

We may therefore build assessments around tasks suggested by an appropriate psychology but determine whether measurement models at appropriate levels of detail provide adequate summaries of evidence. In some applications, we may wish to model competencies with detailed structures suggested by the psychology of learning in the domain. An example is tutoring an individual student's foreign language proficiencies. The ACTFL scale doesn't capture the distinctions we'd need to help a mid-novice become a high-novice. In other applications, such as documenting a student's progress, the ACTFL scale may suffice.

Because measurement model results are, at best, gross summaries of aspects of students' thinking and problem solving abilities, we are obliged to identify contexts that circumscribe their usefulness. Two similar scores convey similar meanings to the extent that they summarize performances on suitably similar tasks, in suitably similar ways, for suitably similar students. We must be alert to patterns in individual students' data that cast doubt on using their test scores to compare them to other students, and we must be reluctant to infer educational implications without examining qualitatively different kinds of evidence.

Two similar scores convey similar meanings to the extent that they summarize performances on suitably similar tasks, in suitably similar ways, for suitably similar students.

Chains of Inferences

A criminal prosecutor must establish the link between observation and conjecture. Seeing Jennifer's car at the Fotomat Thursday night is direct evidence for the claim that her car was there but indirect evidence that Jennifer herself was. It is indirect and less compelling evidence for the conjecture that she was the burglar. Each additional event or inference between an observation and a conjecture adds

[4] An engineering perspective is apropos: "The engineer combines mathematical and physical insight to generate a representation of an unknown process suitable for research, design, and control. A model does not need to mimic the system; it is sufficient to represent the relevant characteristics necessary for the task at hand." (Linse, 1992, pp. 1)

uncertainty. The greater the number of possibilities to account for an observation, the higher the potential that new information would alter our beliefs. For example, Sam's claim that Jennifer's car had been stolen Thursday morning does not, in and of itself, contain much information about whether the car was at the Fotomat, but it does temper our belief that Jennifer herself was.

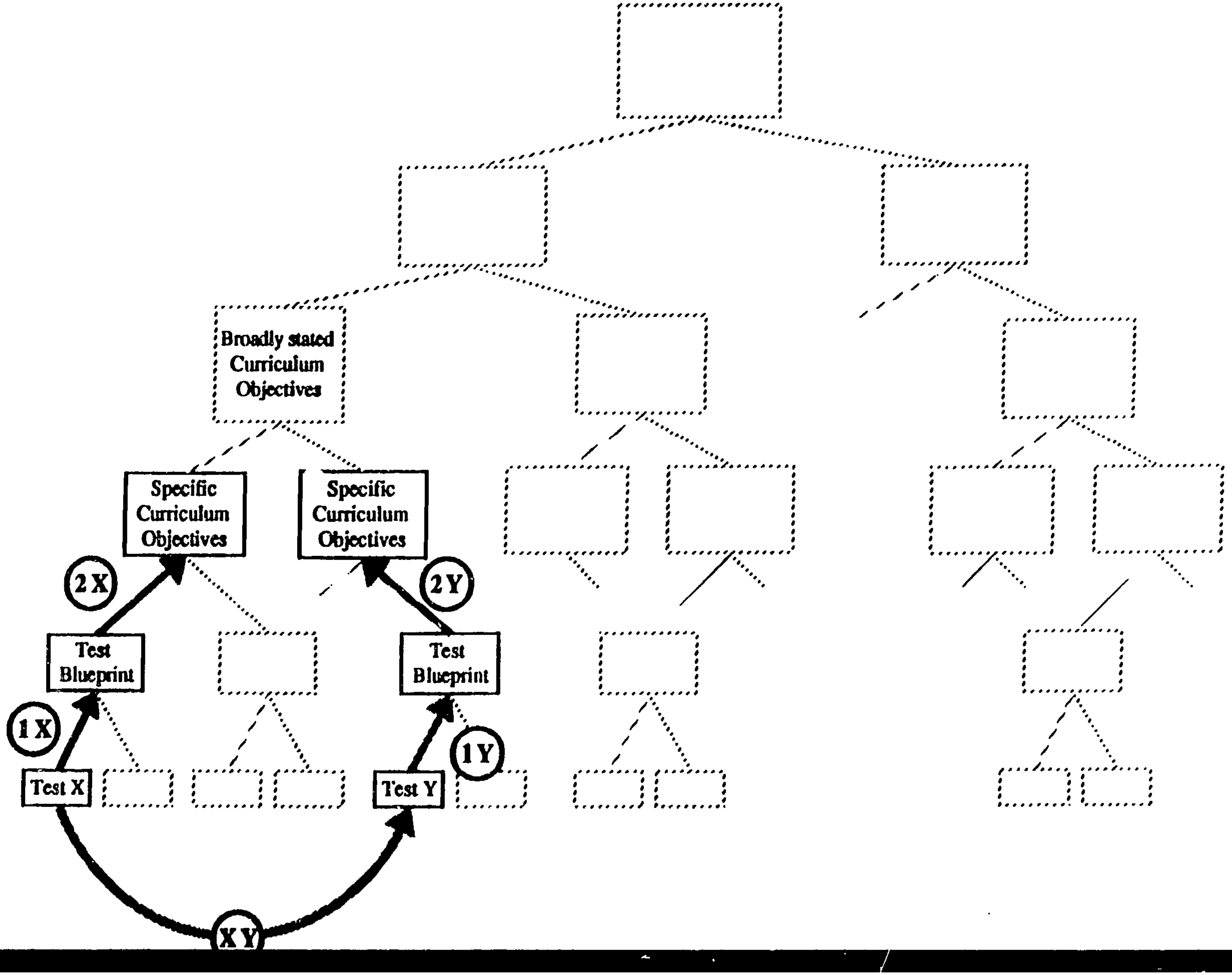
Student behavior in an assessment is closely related to the variables in the conception of competence that generated the assessment. It is less directly related to more abstract levels of competence and even further removed from alternative conceptions of competence that might have led to different assessments. In Figure 2, Test X and Test Y can both be traced back to the same high-level conception of competence, but through different curricular objectives and test specifications. Step 1X represents inference from performance specific to Text X, to general statements at the level of the family of tests sharing the same test specifications. Step 2X represents inference related to the curricular objectives from which the Test X specifications were derived (and from which different test specifications may have also been derived). The chain of inference from Test X to the curricular objectives that led to Test Y involves comparable steps, with an additional "linking" step (Step XY). When we follow this chain, inferences from Test X data to Test Y curricular objectives cannot be more precise than those that would be obtained more directly from Test Y data.

Student behavior in an assessment is closely related to the variables in the conception of competence that generated the assessment. It is less directly related to more abstract levels of competence and even further removed from alternative conceptions of competence that might have led to different assessments.

There are two inferential steps between Test X and its corresponding curricular objectives. For assessment from the behavioral perspective, the first step (Step 1) is direct and can be quite definitive. We observe the frequency of behavior X in a sample of contexts, and draw inferences about the tendency toward behavior X in these kinds of contexts. Subsequent inferences related to a more abstractly defined notion of competence (Step 2) add another layer of uncertainty, however. The relationship between behavioral tendencies in the domain of tasks and the more abstractly defined competence that inspired the task domain can be much less direct. We may be able to estimate behavioral tendencies accurately but face a difficult link in a chain of inference when we want to interpret behavior in terms of competence. Step 1 is comparatively easy; Step 2 is the challenge. In an assessment built from a cognitive perspective, the level of test specifications is more directly related to curricular objectives, but more judgment is required to reason from behavior to the level of test specifications.

Figure 2

Chains of Inference from Test X and Test Y to Two Sets of Specific Curricular Objectives


Broadly stated Curriculum Objectives
Specific Curriculum Objectives
Specific Curriculum Objectives
2 X
2 Y
Test Blueprint
Test Blueprint
1 X
1 Y
Test X
Test Y
X Y

The Roles of Judgment

We now consider some test theory issues concerning judgments. When the potential range of students' performances on an assessment task is constrained to a readily distinguishable set (e.g., multiple-choice options), summarizing performance is straightforward. The tough value judgments have already been made before Kikumi ever sees her test form. These judgments are implicit in the test specifications, which lay out the content and the nature of items, and in the value assignments for each possible response (often simply right or wrong, but possibly previously ordered as to quality or other characteristics). For tasks that allow students to respond with fewer constraints, like writing a letter, performing an experiment, or developing a project over time, judgment is also required *after* the response has been made.

A judge's rating, like the testimony of a witness, is the first link in a chain of inference. The role of test theory is to characterize the uncertainty associated with ratings and guide the construction of a common framework of judgmental standards in order to reduce uncertainty.

This judgment must characterize students' possibly highly individualistic performances within a common frame of reference. The judgment may produce one or more summary statements (e.g., scores, ratings, checklist tallies, evaluative comments); these may be qualitative or quantitative and need not apply to all performances. A judge's rating, like the testimony of a witness, is the first link in a chain of inference. The role of test theory is to characterize the uncertainty associated with ratings and guide the construction of a common framework of judgmental standards in order to reduce uncertainty. Examples and discussions are the best ways to build such a framework—examples that contrast levels of standards that are stated in general terms, feedback on examples analyzed by the group, discussions of actual performances that provoked differences of opinion. This process is analogous to the judgmental ratings of food and pollution Deming mentioned, and the statistical procedures that support them are similar. It can come into play at one or more levels in a given assessment system:

1. *Same conception of competence, same task.* At the most constrained level, judges must be able to provide sufficiently similar ratings of performance for a particular task. For example, students might be asked to write a letter to order a T-shirt from a magazine ad. Judges should be able to agree, to the extent demanded by the purpose, in their ratings of a sample of students' letters. Discrepencies among judges' ratings of the same performance reflect a

source of uncertainty that must be taken into account in inferences about students' competencies.

2. *Same conception of competence, different tasks.* At this level, judges must relate performances on different tasks, or more varied performances within a less standardized setting, to values of a more generally stated competency variable. Students may be asked to write about different topics, for example, or they may be allowed to choose the topic about which they write. Judges should be able to agree, again to an extent demanded by the purpose, about how the different essays relate to the competence variable. This kind of judgment is required for any assessment that personalizes tasks for different students while gathering information about the same generally stated competency. Nested levels can be conceived, as in the ACTFL language assessment guidelines in Table 2. We can study, through statistics and discourse, ratings of diverse performances first within the same language, then from different languages. We will return to this topic in the section on calibration.

3. *Different conceptions of competence, different tasks.* Judges might attempt to relate information from different tasks to different generally described competencies. We might require judgments to determine the weight of evidence that performances designed to inform conjectures in one frame of reference convey for conjectures conceived under a different frame of reference. We shall discuss this further in the sections on projection and moderation.

Assuring fairness, equity, and consistency in applying standards becomes important in high-stakes applications. Developing a statistical framework for such a system serves two functions: (1) as a quality assurance mechanism, to flag unusual or atypical ratings as a safeguard against biases and inconsistent applications of criteria, and (2) as a means of quantifying the typical uncertainties associated with the scoring of students' performances even in the absence of anomalies.

In all these cases, the raters' judgments play a crucial role in assessing students' competence. Assuring fairness, equity, and consistency in applying standards becomes important in high-stakes applications. Developing a statistical framework for such a system serves two functions: (1) as a quality assurance mechanism, to flag unusual or atypical ratings as a safeguard against biases and inconsistent applications of criteria, and (2) as a means of quantifying the typical uncertainties associated with the scoring of students' performances even in the absence of anomalies. Recognizing that any judgment is a unique interaction between the qualities of individual performance and the personal perspective of an individual judge, such a framework enables us to tackle the practical problems that inevitably arise: How do we help judges learn to make these judgments? How do we ascertain the degree of consistency among judges' perceptions of the rating dimensions, and the extent of agreement among their judgments? How do we arrange the numbers and designs of judgments to assure quality for our purposes?

Linking Tests: An Overview

...central problems related to linking two or more assessments are (1) discerning the relationships among the evidence the assessments provide about conjectures of interest, and (2) figuring out how to interpret this evidence correctly.

The central problems related to linking two or more assessments are (1) discerning the relationships among the evidence the assessments provide about conjectures of interest, and (2) figuring out how to interpret this evidence correctly. Deming continues:

> *The number of samples for testing, how to select them, how to calculate estimates, how to calculate and interpret their margin of uncertainty, tests of variance between instruments, between operators, between days, between laboratories, the detection and evaluation of the effect of non-sampling errors, are statistical problems of high order. The difference between two methods of investigation (questionnaire, test) can be measured reliably and economically only by statistical design and calculation.*
>
> W. Edwards Deming,
> 1980, pp. 259

Linking is strongest and simplest if Assessment Y has been constructed from the same blueprint as Assessment X.

Summaries of methods that tackle this problem in educational assessment appear below. Table 3 summarizes their key features. Subsequent sections will describe and illustrate each method in greater detail.

- *Equating.* Linking is strongest and simplest if Assessment Y has been constructed from the same blueprint as Assessment X. Under these carefully controlled circumstances, the weight and nature of evidence the two assessments provide about a broad array of conjectures is practically identical. By matching up score distributions from the same or similar students, we can construct a one-to-one table of correspondence between scores on X and scores on Y to approximate the following property: Any question that could be addressed using X scores can be addressed in exactly the same way with corresponding Y scores, and vice versa.

- *Calibration.* A different kind of linking is possible if Assessment Y has been constructed to provide evidence about the same conception of competence as

Table 3
Methods of Linking Educational Assessments

Link	Description	Procedure	Example	Comments
Equating	Equated scores from tests taken to provide equivalent evidence for all conjectures. Score levels and weights of evidence match up between scores on tests.	1. Construct tests from same blueprint. 2. Estimate distribution of tests in given population. 3. Make correspondence table that matches distributions.	Two forms of a driver's license test, written to the same content and format specifications.	Foundation is not statistical procedure but the way tests are constructed.
Calibration	Tests "measure the same thing," but perhaps with different accuracy or in different ways. Results from each test are mapped to a common variable, matching up the most likely score of a given student on all tests.	Case 1: Use same content, format, and difficulty blueprint to construct tests, but with more or fewer items on different tests. Expected percents correct are calibrated. Case 2: Construct tests from a collection of items that fits an IRT model satisfactorily. Carry out inferences in terms of IRT proficiency variable. Case 3: Obtain judgments of performances on a common, more abstractly defined variable. Verify consistency of judgments (varieties of statistical moderation).	Case 1: A long form and a short form of an interest inventory questionnaire. Case 2: NAEP geometry subscale for grades 4 and 8, connected by IRT scale with common items. Case 3: Judges' ratings of AP Studio Art portfolios including student-selected art projects.	Correspondence table matches up "best estimates," but because weights of evidence may differ, the distribution of "best estimates" can differ over tests. Same expected point estimates for individual students, but with differing accuracy. Different estimates of many group characteristics, e.g., variance and population proportion above cut point.
Projection	Tests don't "measure the same thing," but can estimate the empirical relationships among their scores. After observing score on Y, you can calculate what you'd be likely to observe if X were administered.	Administer tests to the same students and estimate joint distribution. Can derive predictive distribution for Test X performance, given Test Y observation. Can be conditional on additional information about student.	Determine joint distribution among students' multiple-choice science scores, lab notebook ratings, and judgments of observed experimental procedures.	What Test Y tells you about what Test X performance might have been. Can change with additional information about a student. Estimated relationships can vary with the group of students in the linking study and over time in ways that distort trends and group comparisons.

(continued)

Table 3, (continued)
Methods of Linking Educational Assessments

Link	Description	Procedure	Example	Comments
Statistical moderation	Tests don't "measure the same thing," but can match up distributions of their scores in real or hypothetical groups of students to obtain correspondence table of "comparable" scores.	Case 1: If you can administer both X and Y to same students, estimate X and Y distributions. Align X and Y with equating formulas. Case 2: If not, administer X and "moderator" Assessment Z to one group, and Z and Y to another. Impute X and Y distributions for hypothetical common group. Use formulas of equating to align X and Y.	Case 1: Correspondence table between SAT and ACT college entrance exams, based on students who took both. Case 2: Achievement results from History, Spanish, and Chemistry put on "comparable" scales, using common SAT-V and SAT-M tests as moderators.	Comments for projection also apply to statistical moderation. "Comparable" scores need not offer comparable evidence about nature of students' competence. Rather, they are perceived to be of comparable value in a given context, for a given purpose.
Social moderation	Tests don't "measure the same thing," but can match up distributions by direct judgment to obtain correspondence table of "comparable" scores.	Obtain samples of performances from two assessments. Have judges determine which levels of performance on the two are to be treated as comparable. Can be aided by performance on a common assessment.	Obtain samples of Oregon and Arizona essays, each rated through their own rubrics. Determine, through comparisons of scores given to examples, "comparable" levels of score scales.	"Comparable" scores need not offer comparable evidence about nature of students' competence. They are perceived to be of comparable value in a given context, for a given purpose.

Assessment X, but the kinds or amounts of evidence differ. The psychometric meaning of the term “calibration” is analogous to its physical measurement meaning: Scales are adjusted so that the expected score of a student is the same on all tests that are appropriate to administer to him or her.[5] Unlike equating, which matches tests to one another directly, calibration relates the results of different assessments to a common frame of reference, and thus to one another only indirectly. Some properties of calibration are disconcerting to those familiar only with equating: As a consequence of different weights of evidence in X and Y data, the procedures needed to give the right answers to some X questions from Y data give the wrong answers to others. It is possible to answer X questions with Y data if a calibration model is suitable, but generally *not* by means of a single correspondence table. We discuss three settings in which calibration applies: (1) constructing tests of differing lengths from essentially the same blueprint, (2) using item response theory to link responses to a collection of items built to measure the same construct, and (3) soliciting judgments in terms of abstractly defined criteria.

Unlike equating, which matches tests to one another directly, calibration relates the results of different assessments to a common frame of reference, and thus to one another only indirectly.

- *Projection.* If assessments are constructed around different types of tasks, administered under different conditions, or used for purposes that bear different implications for students’ affect and motivation, then mechanically applying equating or calibration formulas can prove seriously misleading: X and Y do not “measure the same thing.” This is not merely a matter of stronger or weaker information but of qualitatively different information. If it is sensible to administer both X and Y to any student, statistical machinery exists to (1) estimate, in a linking study, relationships among scores from X and Y, and other variables in a population of interest and then (2) derive *projections* from Y data about what the answers to the X questions might have been, in terms of a probability distribution for our expectations about the possible outcomes. As X and Y become increasingly discrepant—such as when they are meant to provide evidence for increasingly different conceptions of competence—the evidential value of Y data for X questions drops, and projections

If it is sensible to administer both X and Y to any student, statistical machinery exists to derive *projections* from Y data about what answer to X questions might have been.

[5] Some writers use the terms “equating’ and “calibrating” interchangeably to describe what I call calibrating. The differences in procedures and properties I describe here merit maintaining the distinction.

become increasingly sensitive to other sources of information. The relationship between X and Y can differ among groups of students and can change over time in response to policy and instruction.

- *Moderation*. Certain assessment systems obtain Test X scores from some students and Test Y scores from others, under circumstances in which it isn't sensible to administer both tests to any student. Literature students take a literature test, for example, and history students take a history test. There is no pretense that the two tests measure the same thing, but scores that are in some sense comparable are desired nevertheless. Whereas projection evaluates the evidence that results on one assessment provide about likely outcomes on another, *moderation* simply aligns scores from the two as to some measure of comparable worth. The way that comparable worth is determined distinguishes two varieties of moderation:

Whereas projection evaluates the evidence that results on one assessment provide about likely outcomes on another, *moderation* simply aligns scores from the two as to some measure of comparable worth.

1) *Statistical moderation* aligns X and Y score distributions, sometimes as a function of joint score distributions with a third "moderator test" that all students take. That is, a score on X and a score on Y are deemed comparable if the same proportion of students in a designated reference population (real or hypothetical) attains scores at or above those two levels on their respective tests. The score levels that end up matched can depend materially on the choice of the reference population and, if there is one, the moderator test. Historically, these procedures have been discussed as a form of "scaling" (Angoff, 1984).

2) *Social moderation* uses judgment to match levels of performance on different assessments directly to one another (Wilson, 1992). This contrasts with the judgmental linking discussed under calibration, which relates performances from different assessments to a common, more abstractly defined variable, and under projection, which evaluates the evidence that judgmental ratings obtained in one context hold for another context.

Equating, calibration, and moderation address the correspondence between single scores from two assessments. Two assessments can each have multiple scores, however, and these approaches could apply to matched sets of scores. Projection can address joint relationships among all scores from multiple assessments simultaneously.

A Note on Linking Studies

All of the linking methods described here require data of one kind or another, such as student responses or judges' ratings. If the data are collected in a special linking study outside the normal data-gathering context of any of the assessments, then the possible consequences of differences in student or rater behavior between the linking study and the usual context constitute an additional link of inference from one assessment to the other. One must take care to minimize the uncertainty this step engenders, and, whenever possible, quantify its effect on inferences (e.g., Mislevy, 1990). This caveat extends to psychological as well as physical conditions. Differences in levels of individual performance can easily arise between high-stakes and low-stakes administrations of the same tasks.

Differences in levels of individual performance can easily arise between high-stakes and low-stakes administrations of the same tasks.

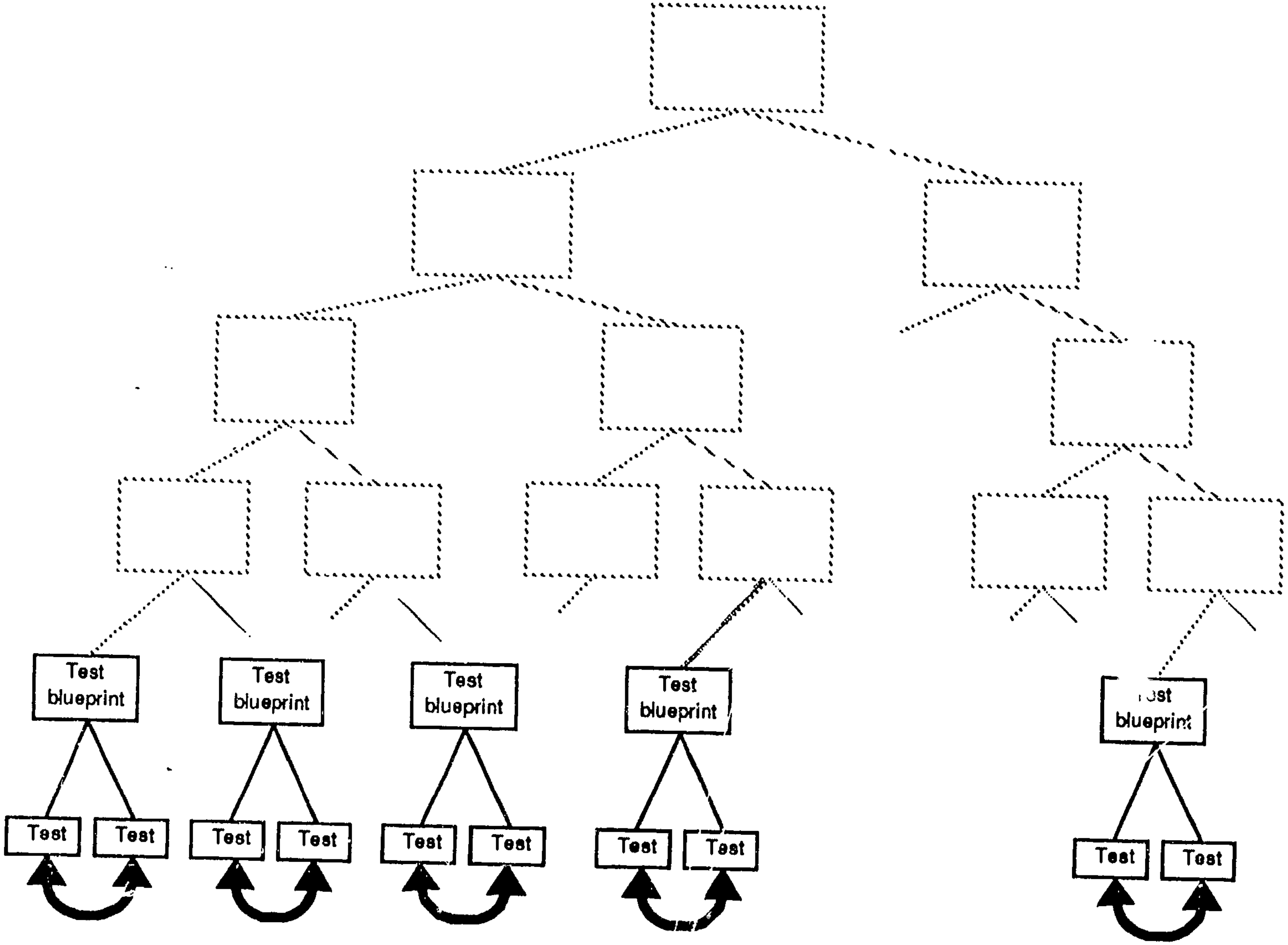
Equating

Selection and placement testing programs update their tests periodically, as the content of specific items becomes either obsolete or familiar to prospective examinees. New test forms are constructed according to the same blueprint as previous forms, with the same number of items asking similar questions about similar topics in the same ways. SAT Mathematics test developers, for example, maintain a balance among items with no diagrams, with diagrams drawn to scale, and diagrams not drawn to scale, along with a hundred other formal and informal constraints.[6] Figure 3 relates an equating link to the hierarchy of competence definitions introduced earlier. Pretest samples of examinee responses to the new items are gathered along with responses to items from previous forms, and items' statistical properties from pretest samples may also be used to select items for new forms. The objective is to create "parallel" test forms, which provide approximately equivalent evidence for a broad range

[6] Stocking, Swanson, & Pearlman (1991) describe how these constraints were specified and employed in a demonstration of an automated test assembly algorithm.

Figure 3

Equating Linkages



of potential conjectures. *Equating* makes slight adjustments in the results of such test forms (which were expressly constructed to make such adjustments negligible!) by aligning the distributions of scores from the same or similar students on the two forms.

Equating in Physical Measurement

Two hundred years ago, Karl Friedrich Gauss studied how to estimate the "true position" of a star from multiple observations—all similar but not identical because of the inherent imperfections of the telescope-and-observer system. If each observation differs from the true value by a "measurement error" unrelated to either the true value or the errors of other observations, then (1) the average of the measurements is a good estimate of the true value, and (2) the estimate can be made more precise by averaging over more observations. The nature of the measuring instrument determines the typical size and distribution of the measurement errors. If two instruments react to exactly the same physical property with exactly the same sensitivity, their readings can be *equated,* in the following sense: A table of correspondence can be constructed so that a value from one instrument has exactly the same interpretation and measurement error distribution as the corresponding value would from the other instrument.

Equating Temperature Readings from Two Similar Thermometers

Table 4 gives July temperatures for the 60 U.S. cities.[7] We will use these as "true scores," from which to construct "observed scores" from some hypothetical measuring instruments. The top panel in Figure 4 shows the distribution of true temperatures, highlighting the nine cities from the southeast and six from the west coast; Figure 5 shows the locations of the cities. In analogy to achievement testing, however, we never observe these true measures directly. Instead, we observe readings from two "noisy" measures, Thermometer X and Thermometer Y; that is, each reading may be a bit higher or lower than the true temperature.

[7] These data are taken from the SMSAs demonstration data set from *Data Desk Professional 2.0* statistical package (Velleman, 1988).

Figure 4

Distributions of True Temperatures and Temperature Readings

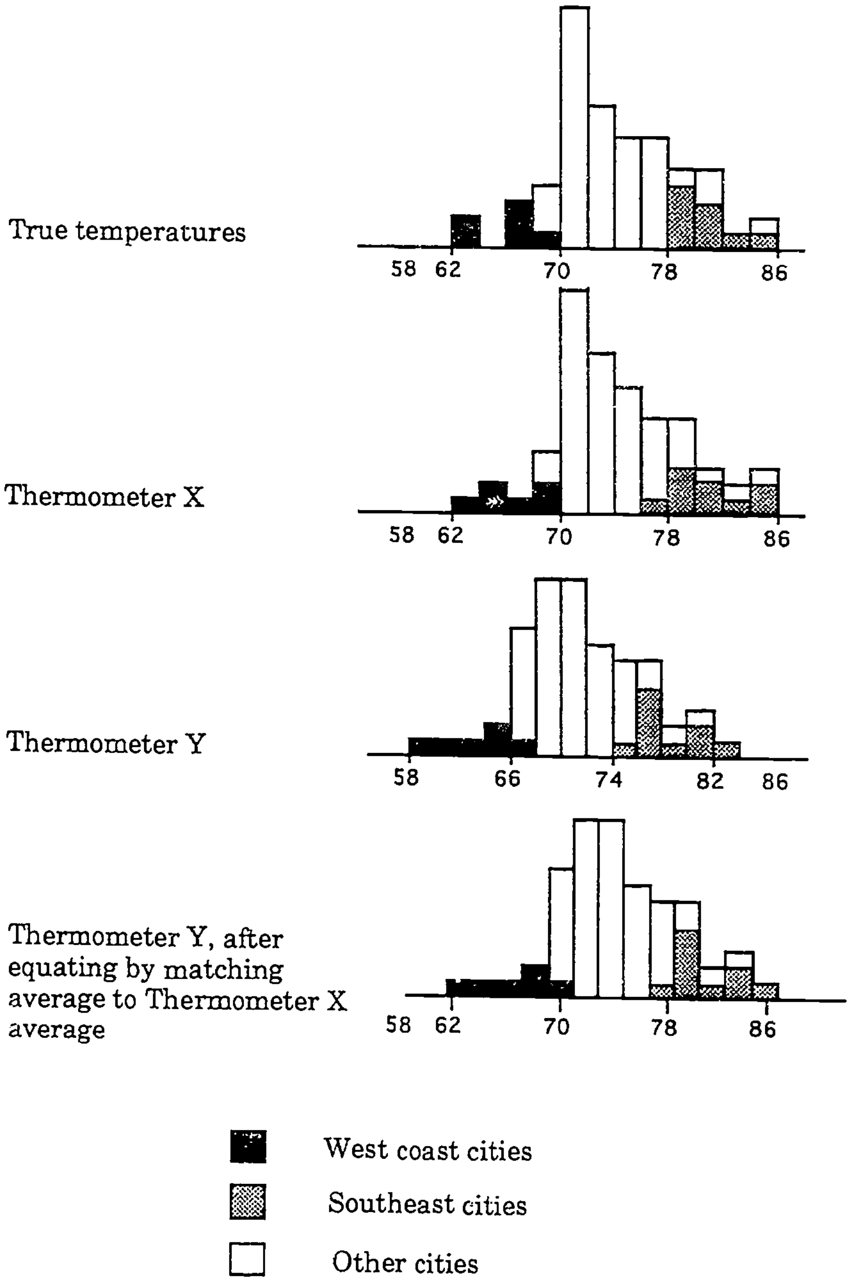


Table 4
Temperature Measures for 60 Cities

			Thermometers				
City	Latitude	"True" Temp	X	Y	Equated Y	Cricket	Equated Cricket
NORTH/CENTRAL							
Akron, OH	41.1	71	70.3	68.3	71.1	61.3	64.1
Albany, NY	42.4	72	73.5	69.6	72.4	68.2	70.8
Allentown, PA	40.4	74	71.8	71.3	74.1	74.4	73.3
Baltimore, MD	39.2	77	75.2	73.1	75.9	76.2	75.2
Boston, MA	42.2	74	73.6	72.1	74.9	67.9	70.3
Bridgeport, CT	41.1	73	73.3	70.4	73.2	82.2	78.6
Buffalo, NY	42.5	70	70.2	67.6	70.4	72.8	73.0
Canton, OH	40.5	72	71.3	69.7	72.5	72.7	72.7
Chicago, IL	41.5	76	75.8	72.2	75.0	79.4	77.2
Cincinnati, OH	39.1	77	78.0	74.5	77.3	86.3	84.2
Cleveland, OH	41.3	71	68.9	67.8	70.6	71.5	71.7
Columbus, OH	40.0	75	75.5	71.6	74.4	84.2	79.6
Dayton, OH	39.5	75	75.3	71.8	74.6	75.5	74.5
Denver, CO	39.4	73	73.7	70.9	73.7	80.4	77.4
Detroit, MI	42.1	74	73.3	71.5	74.3	66.4	70.0
Flint, MI	43.0	72	71.9	67.8	70.6	76.5	75.3
Fort Worth, TX	32.5	85	85.2	81.8	84.6	83.8	79.4
Grand Rapids, MI	43.0	72	71.6	69.1	71.9	71.9	72.0
Greensboro, NC	36.0	77	76.5	75.1	77.9	77.3	75.9
Hartford, CT	41.5	72	72.3	68.0	70.8	73.2	73.3
Indianapolis, IN	39.5	75	74.7	70.8	73.6	84.6	80.2
Kansas City, MO	39.1	81	81.0	77.3	80.1	75.5	74.7
Lancaster, PA	40.1	74	74.5	71.6	74.4	75.2	73.7
Louisville, KY	38.2	71	70.9	67.1	69.9	71.0	71.6
Milwaukee, WI	43.0	69	70.4	66.2	69.0	71.9	71.9
Minneapolis, MN	44.6	73	70.8	69.5	72.3	70.8	71.2
New Haven, CT	41.2	72	72.0	70.8	73.6	71.9	72.3
New York, NY	40.4	77	77.0	74.0	76.8	77.3	76.5
Philadelphia, PA	40.0	76	74.9	72.7	75.5	81.3	78.0
Pittsburgh, PA	40.3	72	72.7	69.4	72.2	71.6	71.8

(continued)

Table 4, continued
Temperature Measures for 60 Cities

City	Latitude	"True" Temp	Thermometers X	Thermometers Y	Equated Y	Cricket	Equated Cricket
Providence, RI	41.5	72	71.9	69.0	71.8	63.7	67.8
Reading, PA	40.2	77	77.2	73.1	75.9	71.8	71.9
Richmond, VA	37.4	78	77.4	75.6	78.4	76.0	74.9
Rochester, NY	43.2	72	71.2	68.6	71.4	85.0	80.6
St. Louis, MO	38.4	79	79.0	76.2	79.0	85.4	82.0
Springfield, MA	42.1	74	73.0	70.2	73.0	77.3	75.8
Syracuse, NY	43.1	72	71.7	68.0	70.8	64.5	68.2
Toledo, OH	41.4	73	73.3	71.7	74.5	79.2	77.0
Utica, NY	43.1	71	70.1	69.5	72.3	65.3	68.9
Washington, DC	38.5	78	78.3	74.7	77.5	68.1	70.4
Wichita, KS	37.4	81	82.0	79.2	82.0	74.9	73.6
Wilmington, DE	39.5	76	75.9	72.9	75.7	72.1	72.7
Worcester, MA	42.2	70	69.7	66.3	69.1	74.6	73.3
York, PA	40.0	76	76.7	73.3	76.1	67.9	70.2
Youngstown, OH	41.1	72	72.7	67.7	70.5	70.9	71.3
SOUTHEAST							
Atlanta, GA	33.5	79	77.0	77.3	80.1	86.2	82.2
Birmingham, AL	33.3	80	78.6	77.2	80.0	86.9	84.7
Chattanooga, TN	35.0	79	80.2	75.9	78.7	82.2	78.3
Dallas, TX	32.5	85	84.7	82.0	84.8	77.1	75.5
Houston, TX	29.5	84	84.2	80.4	83.2	79.0	77.0
Memphis, TN	35.1	82	82.2	78.0	80.8	85.1	81.0
Miami, FL	25.5	82	80.6	80.4	83.2	82.7	79.0
Nashville, TN	36.1	80	79.6	76.0	78.8	88.3	85.2
New Orleans, LA	30.0	81	79.4	76.8	79.6	78.9	76.7
WESTCOAST							
Los Angeles, CA	34.0	68	65.9	66.0	68.8	61.6	65.6
Portland, OR	45.3	67	67.8	62.5	65.3	65.6	69.7
San Diego, CA	32.4	70	70.0	65.4	68.2	74.6	73.5
San Francisco, CA	37.5	63	63.6	59.8	62.6	55.8	63.6
San Jose, CA	37.2	68	68.2	65.8	68.6	70.5	70.9
Seattle, WA	47.4	64	64.1	61.7	64.5	67.5	70.1

We've had Thermometer X for some time and determined that it is about right on the average.[8] The second panel in Figure 5 shows the distribution of its measures. The average over cities is about the same as the true average, and the average for the regions is about right. About the same number of cities are too high and too low, compared to their true values. On the whole, the Thermometer X readings are spread out a little more than the true values, because the measurement errors add variance to the collection of measures. (We shall see how this point becomes important in *calibration*.)

We now acquire Thermometer Y, the same kind of thermometer as X, with the same accuracy. It comes from the factory with an adjustment knob to raise or lower all its readings—which, unfortunately, doesn't work. We need to make a table to translate Thermometer Y readings so they match Thermometer X readings. The unadjusted readings are shown in the third panel of Figure 5 and plotted against Thermometer X readings in Figure 6.[9] The spread of readings is about the same for both, but the values from Y are systematically lower. We find the average difference over all 60 cities: 2.8. Adding 2.8 to Thermometer Y readings gives them the same average as the Thermometer X readings. This "mean equating" matches up the two thermometers pretty well on the average, yielding the distribution in the last panel of Figure 5. The equating is operationalized in terms of the correspondence table shown as Table 5.

Equating in Educational Assessment

Edgeworth (1888, 1892) and Spearman (1904, 1907) launched classical test theory (CTT) around the turn of the century by applying the ideas of true-score measurement to tests. CTT views the average (or, equivalently, the total) of 1-for-right/0-for-wrong results from numerous test items as an imperfect measure of an examinee's "true score." While each

[8] To produce Thermometer X readings, I added to each "true" temperature a number drawn at random from a normal distribution, with mean zero and standard deviation one.

[9] Thermometer Y readings are "true" temperature, minus 3, plus a random number from a normal distribution, with mean zero and standard deviation one.

Figure 5
Sixty Cities

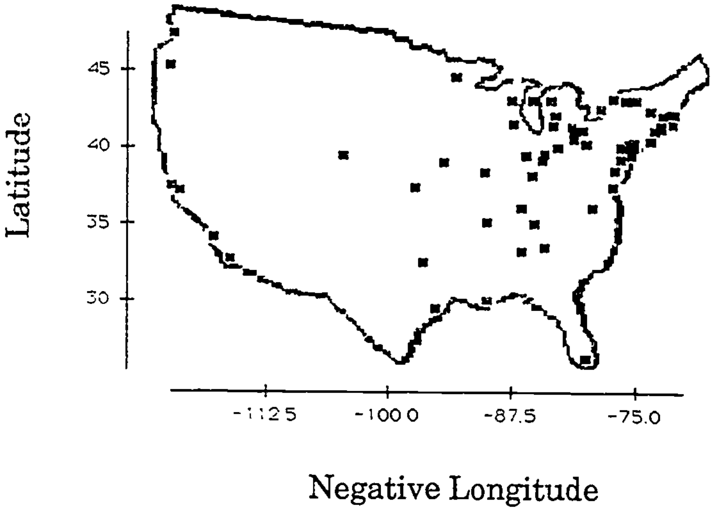


Figure 6

Unadjusted Thermometer Y Readings Plotted Against Thermometer X Readings

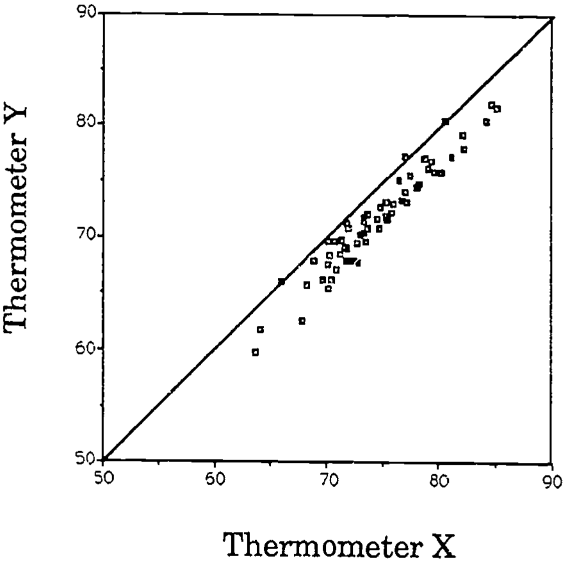


Table 5
Correspondence Table for Thermometer X and Thermometer Y

Thermometer X Reading	Thermometer Y Reading
⋮	⋮
85.8	83.0
85.7	82.9
85.6	82.8
85.5	82.7
85.4	82.6
85.3	82.5
85.2	82.4
85.1	82.3
85.0	82.2
84.9	82.1
84.8	82.0
84.7	81.9
84.6	81.8
84.5	81.7
84.4	81.6
84.3	81.5
84.2	81.4
84.1	81.3
84.0	81.2
83.9	81.1
83.8	81.0
83.7	80.9
83.6	80.8
83.5	80.7
83.4	80.6
83.3	80.5
83.2	80.4
83.1	80.3
83.0	80.2
82.9	80.1
82.8	80.0
82.7	79.9
82.6	79.8
82.5	79.7
82.4	79.6
82.3	79.5
82.2	79.4
⋮	⋮

of the separate items taps specific skills and knowledge, a score from such a test captures a broad general tendency to get items correct and provides information for conjectures about competencies also defined broadly (Green, 1978). Different tests drawn from the same domain correspond to repeated measures of the same true score.

Equating applies to tests that embody not merely the same general statement of competence but structurally equivalent operational definitions. But even with the care taken to create parallel test forms, scores may tend to be a bit higher or lower on the average with a new form than with the previous one. Equating adjusts for these differences.

Equating applies to tests that embody not merely the same general statement of competence but structurally equivalent operational definitions. But even with the care taken to create parallel test forms, scores may tend to be a bit higher or lower on the average with a new form than with the previous one. The new form's scores may spread examinees out a little more or less. An equating procedure adjusts for these overall differences, so that knowing which particular form of a test an examinee takes no longer conveys information about the score we'd expect. There are a variety of equating schemes (see Petersen, Kolen, & Hoover, 1989), but the "equivalent groups" design captures the essence:

1. Administer Test X and Test Y to randomly selected students from the same specified group—typically, a group representative of students with whom the tests will be used.

2. Make a correspondence table matching the score on Test X that 99% of the sample was above to the score on Test Y that 99% of that sample was above. Do the same with the 98% score level, the 97% level, and so on. If the X and Y distributions have similar shapes, we may be able to align the distributions sufficiently well by just matching up the means of the X and Y samples, as we did in the temperature example, or the means and standard deviations, as in "linear equating."

3. Once a correspondence table has been drawn up, look up any score on Test Y, transform it to the Test X score on the table, and use the result exactly as if were an observed Test X score with that value. The same table can be used in the same way to go from X scores to equated Y scores.

Long tests constructed from a given domain of tasks correspond to more accurate measures than short tests drawn from the same domain. Equating, strictly defined, applies to two similarly constructed long tests, or two similarly constructed short ones. The calibration section illustrates the problems that arise if we use equating procedures to link a long test and a short one.

The indicator of a test's accuracy under CTT is *reliability*, a number between 0 and 1 gauging the extent of agreement between different forms of the test. This number depends not only on the precision with which the test measures students' true scores, but also on the *amount of variation among true scores* in the student population. This definition reflects the classic norm-referenced usage of tests: lining up examinees along a single dimension for selection and placement. A test that measures examinees' true scores perfectly accurately is useless for this purpose if all their true scores are the same, so this test has a reliability of 0 for this population—although it might separate examinees in a different population quite well and have a reliability closer to one for that group. The plot of Thermometer X and Thermometer Y readings shown in Figure 6 corresponds to a reliability of .96 for the sample as a whole—higher than the reliability of the two-hour long SAT. The reliability of the thermometer readings is .85 for the cities in the southeast only.

...reliability depends not only on the precision with which the test measures students' true scores, but also on the *amount of variation among true scores* in the student population. This definition reflects the classic norm-referenced usage of tests: lining up examinees along a single dimension for selection and placement.

A high reliability coefficient is important when scores are used for high-stakes, norm-referenced uses with individual students, because high reliability means that a different sample of items of the same kind would lead to the same decision about most students. This same index can be less important—even inappropriate—for other purposes. A shorter and less reliable test can suffice for less consequential norm-referenced decisions—a quiz to determine whether to move on to the next chapter, for example. When the purposes of assessment are criterion-referenced, evidence about the accuracy with which an individual's competencies are assessed is more important than evidence about how he or she compares with other students. And accurate estimates of aspects of the distribution of a group of students' true scores can be obtained from short tests, even if measurement for individuals is quite inaccurate (Lord, 1962). NAEP, which was designed to provide information at the level of groups rather than individuals, thus administer a small sample of items from a large pool to each student, thereby amassing substantial evidence about groups on a broader range of tasks than could be administered to any single student.

A high reliability coefficient is important when scores are used for high-stakes, norm-referenced uses with individual students, because high reliability means that a different sample of items of the same kind would lead to the same decision about most students.

Test construction and equating are inseparable.

Reliability is a sensible summary of the evidence a test provides *in a specific context* (a particular group of students), *for a specific purpose* (determining how those students line up in relation to one another). It does *not* indicate the strength of evidence for statements about the competencies of individual examinees. In the section on calibration, we discuss item response theory indices of evidence that are defined more along these lines.

Comments on Equating

When equating does work, it works because of the way the *tests are constructed*, not simply because of the way the linking data are collected or correspondence tables built.

Test construction and equating are inseparable. When they are applied in concert, equated scores from parallel test forms provide virtually exchangeable evidence about students' behavior on the same general domain of tasks, under the same specified standard conditions. When equating does work, it works because of the way the *tests are constructed*, not simply because of the way the linking data are collected or correspondence tables built. Tests constructed in this way can be equated whether or not they are actually measuring something in the deeper senses discussed in the next section—although of course determining what they measure, if indeed anything, is crucial for justifying their use.

Calibration

Calibration relates observed performance on different assessments to a common frame of reference. Properly calibrated instruments have the same expected value for measuring an object with a given true value.

Calibration relates observed performance on different assessments to a common frame of reference. Properly calibrated instruments have the same expected value for measuring an object with a given true value. These instruments provide evidence about the same underlying variable, but, in contrast to equating, possibly in different amounts or by different means or in different ranges. A yardstick is less accurate than vernier calipers but is cheaper and easier to use for measuring floor tiles. Nevertheless, the yardstick and calipers should give us about the same measurement for a two-inch bolt. The focus is on inference from the measuring instrument to the underlying variable, so calibrated instruments are related to one another only indirectly. This too contrasts with equating, where the focus is on matching up observed performance levels on tests to one another directly. After illustrating properties of calibration in the context of temperature, we discuss three cases of calibration in educational assessment.

Calibrating a "Noisy" Measure of Temperature

The warmer it gets, the faster the snowy tree cricket (*Oecanthus fultoni*) chirps. The relationship between temperature and cricket chirps isn't as strong as the relationship between temperature and the density of mercury, and it isn't very useful below freezing or above about 100°. Calibrating a cricket—let's call him Jiminy—involves transforming the counts of his chirps so that for any true temperature, the resulting estimate of temperature is as likely to be above the true value as below it. The *Encyclopedia Brittanica* gives "the number of chirps in 15 seconds + 40" as an approximation for Fahrenheit temperature.

If the true temperature is 65°, five calibrated temperature readings from Jiminy's chirps could be 57°, 61°, 66°, 68°, and 71°—averaging around the true value but quite spread out. In contrast, five readings from Thermometer X could give readings of 63°, 64°, 65°, 65°, and 67°—also averaging around the true value, but with much less spread. If we have a reading of 80° from Thermometer X or Jiminy, our best estimate of the true temperature is 80° either way, but the evidence that the true temperature is around 80° rather than 75° or 85° is much stronger with Thermometer X.

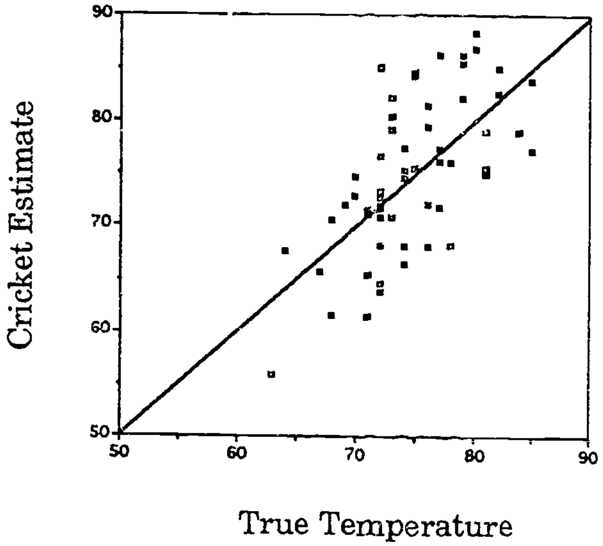
"Calibrated" cricket measures give *unbiased* answers—each one possibly different from the true value but tending to produce the right answer on the average—to questions about the temperature of an individual city or about the average of a group of cities.

Table 4 contains a column with a hypothetical set of readings from Jiminy's calibrated measures, and Figure 7 plots them against true temperature around the country.[10] These calibrated cricket readings correspond to the true temperatures most likely to have produced them.[11] They give *unbiased* answers—each one possibly different from the true value but tending to produce the right answer on the average—to questions about the temperature of an individual city or about the average of a group of cities:

[10] Jiminy's readings are "true" temperatures plus a random number from a normal distribution, with mean zero and standard deviation five.

[11] In statistical terms, they are, under plausible assumptions, maximum likelihood estimates. The likelihood function induced by a chirp count is more spread out than that of a Thermometer X or Y reading.

Figure 7
Cricket Temperature Readings
Plotted Against True Temperatures



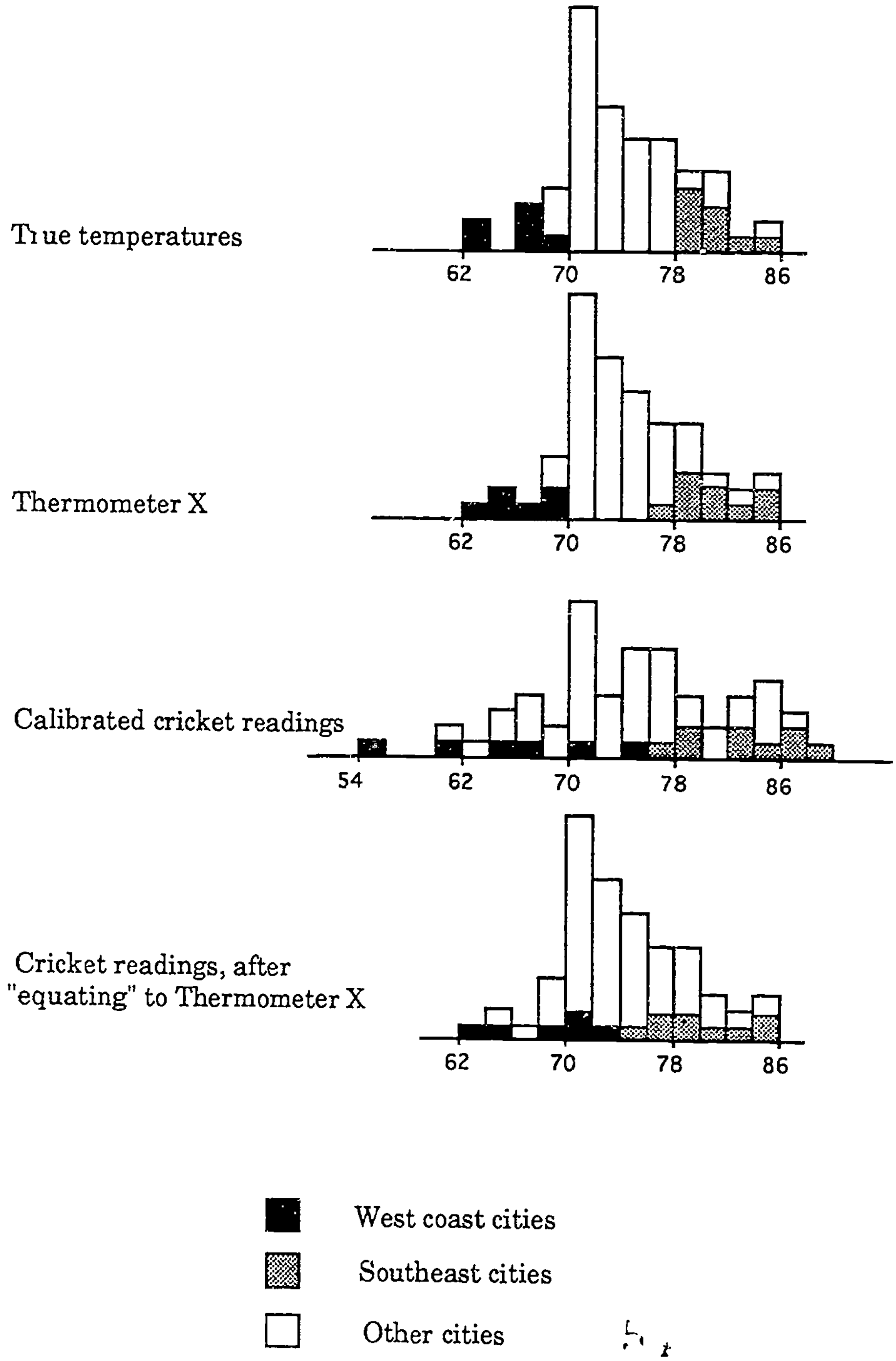
- Is the true temperature of Fort Worth above or below 80°? See whether its calibrated-cricket measure is above or below. It's above. The true value is above, too. On the other hand, N9w Orleans has a true temperature just above 80° but a cricket measure just below 80°. The noisier a measure is, the more errors of this type occur. If they are properly calibrated, though, similar numbers of readings will be too high and too low.

- What are the averages for the entire sample, for the southeast, and the west coast? They turn out pretty close in this example: for the entire sample, 74.9° from the cricket vs. 74.6° actual; 65.9° vs. 66.7° for the west coast; 82.9° vs. 81.3° for the southeast.

Figure 8 compares the distribution of Jiminy's measures to the true temperatures. Even though every cricket measure has the true measure as its expected value, the added variation from the measurement error spreads the set of measures out considerably. This uncertainty is reflected in the

Figure 8

Distributions of True Temperatures and Selected Estimates



spread of cricket measures' likelihood functions, but it is ignored when we use only the best point estimate of the temperature. As a consequence, properly calibrated cricket measures give *biased* answers (tending to have a *wrong* answer as their average) to questions about features of the distribution other than individual and group averages:

- What is the range of temperatures? The true range runs from a low of 63° (San Francisco) to a high of 85° (Dallas and Fort Worth). The calibrated cricket measures run from 55.8° (again San Francisco, but with a low error term added to an already low true value), up to 88.3° (Nashville, with a true temperature of 80° but a high measurement error term).

- What proportion of cities is below 70°, and what proportion is above 80°? The answers for *true* temperatures are 10% and 13%; the calibrated-cricket answers are 22% and 25%.

Could we solve this problem by "equating" the cricket values to match the spread of the true measures? We can, in fact, match cricket measures to the Thermometer X distribution exactly, by mapping the highest cricket value to the highest thermometer value, the next highest to the next highest, and so on, disregarding which cities these values are associated with. The numbers appear in Table 4, and the distribution is shown in the last graþn in Figure 8. Because the way we constructed it, this graph has exactly the same silhouette as that of Thermometer X readings—the same mean and spread, the same proportions of cities below 70° and above 80°, and so on. It would seem that we have made cricket measures equivalent to Thermometer X measures.

But wait! See how the Pacific coast and southeastern cities have shifted toward the center of the distribution. Forcing the distribution to be right for the population as a whole has compromised the cardinal property of calibration, matching up most likely best estimates for individuals. A city truly above 80° is now more than likely to have an "equated" temperature closer to the overall average, or below 80°. About the right proportion of equated measures are indeed up there, but now some of the cities that are above 80° happen to have high measurement error terms. The southeastern cities shrink toward the center of the distribution, their average dropping from the nearly correct 82.9° down to 79°. The Pacific coast cities similarly rise spuriously. Inappropriate equating has thus distorted criterion-referenced inferences about individual cities.

It is possible to estimate, from properly calibrated measures, characteristics of a distribution such as its spread and proportions above selected points. More complex statistical methods are required, however, because one must account for the spread of evidence about each of the observations, not just their best single-point estimates. To estimate the proportion of true values above 70°, for example, requires calculating and adding the probabilities that each of the cities is above 70°—a negligible probability for a city with an accurate Thermometer X reading of 65°, but about a one-out-of-five chance for a city with a noisy cricket measure of 65°. A paradox arises that worsens as the accuracy of measurement declines: The proportion of cities whose best estimate is over 70° is not generally equal to the best estimate of the proportion of cities with true temperatures over 70°. No single correspondence table can resolves the paradox. From these observations, we must use different statistical techniques to extract evidence about different kinds of conjectures.

It is possible to estimate, from properly calibrated measures, characteristics of a distribution such as its spread and proportions above selected points. More complex statistical methods are required, however, because one must account for the spread of evidence about each of the observations, not just their best single-point estimates.

Calibration in Educational Assessment, Case 1: Long and Short Tests Built to the Same Specifications

The simplest calibration situation in educational assessment differs from the equating situation in just one respect: Two tests can be built to the same specifications and use similar items and standard administration conditions, but contain more or fewer items. For example, Test X could consist of three items from each category in a detailed table of specifications, while Test Y had one from each category. Test X would be used to gather information for inferences about individual students, while Test Y could be used for inferences about group averages.

The simplest calibration situation in educational assessment differs from the equating situation in just one respect: Two tests can be built to the same specifications and use similar items and standard administration conditions, but contain more or fewer items.

Bob Linn (in press) uses the example of a basketball free-throw competition to illustrate calibration of varying-length tests. The coach wants to select players who shoot with at least 75% accuracy. The long test form is 20 tries; the short form is four tries. If a player's true accuracy is, say, 50%, her most likely outcome in either setting is 50%: 10 of 20, or two of four. Accordingly, the properly calibrated estimate of the accuracy of a player who makes 50% of her shots is 50%. But because of the greater uncertainty associated with observing only four tries, a true 50% shooter has a probability of .31 of making at least 75% of her shots with the short form, but a probability of less than .01 with the long form.

In the same way, the observed percent-correct scores from two educational tests constructed in this manner are approximately calibrated in terms of expected percent correct on a hypothetical, infinitely long test from the same specifications. (The item response theory methods discussed later could be used to fine-tune the calibration.) Unbiased estimates are obtained for individuals, but the short test would show a wider spread of scores than the long test. The short form would yield more errors about who was or was not above a given true-score level, but this phenomenon is inherent in its sparser evidence and cannot be eliminated through calibration or equating. As with the cricket measures and the Thermometer X readings, building an equating correspondence table for Test X and Test Y would solve this problem, but at the cost of introducing biases in scores for individuals. Inappropriately "equated" scores for the short test would tend to give people scores too close to the average, compared to their true scores.

Properly calibrated scores, then, are appropriate for obtaining an unbiased estimate for each student. They would be the choice for criterion-referenced inferences for individuals. We could construct one-to-one correspondence tables for this purpose for different length tests from the same specifications. But, as with cricket temperatures, the same tables would misestimate the shape of the distribution for a *group* of students, to a degree and in a way that depends on the accuracy of those estimates. For example, shorter tests generally cause additional overestimation of a distribution's spread and more greatly exaggerated estimates of students in the extremes. As we mentioned, correctly estimating the shape of a distribution requires accounting for the uncertainty associated with each observation. Methods for accomplishing this within classical test theory have been available for some time (e.g., Gulliksen, 1950/1987). Analogous methods have appeared more recently for problems in which different observations have different amounts and shapes of uncertainty, as in the item response theory context discussed below (e.g., Mislevy, 1984, 1991).

Calibration in Educational Assessment, Case 2: Item Response Theory for Patterns of Behavior

Standard equating and simple calibration depend on the construction of highly constrained observational settings.

Standard equating and the simple calibration situation described above depend on the construction of highly constrained observational settings. Inferences about behavior in two settings can be related because the settings are so similar. This approach offers little guidance for tests built to different blueprints but intended to measure the same construct, such as harder or easier collections of similar items. Item response theory (IRT) lays out a framework that the interactions of students and items must exhibit to satisfy the axioms of "measurement" as it developed in the physical sciences (see especially Rasch, 1960/1980). We illustrate these ideas in the context of right/wrong items, but extensions to ordered categories, counts, ratings, and other forms of data are available. *If* an IRT model is an adequate description of the patterns that occur in item responses, statistical machinery in the IRT framework enables one to calibrate tests in essentially the same sense as we calibrate physical instruments. In this case, the items and tests constructed from them are calibrated to the unobservable variable "tendency to get items of this kind right."

If an IRT model is an adequate description of the patterns that occur in item responses, statistical machinery in the IRT framework enables one to calibrate tests in essentially the same sense as we calibrate physical instruments.

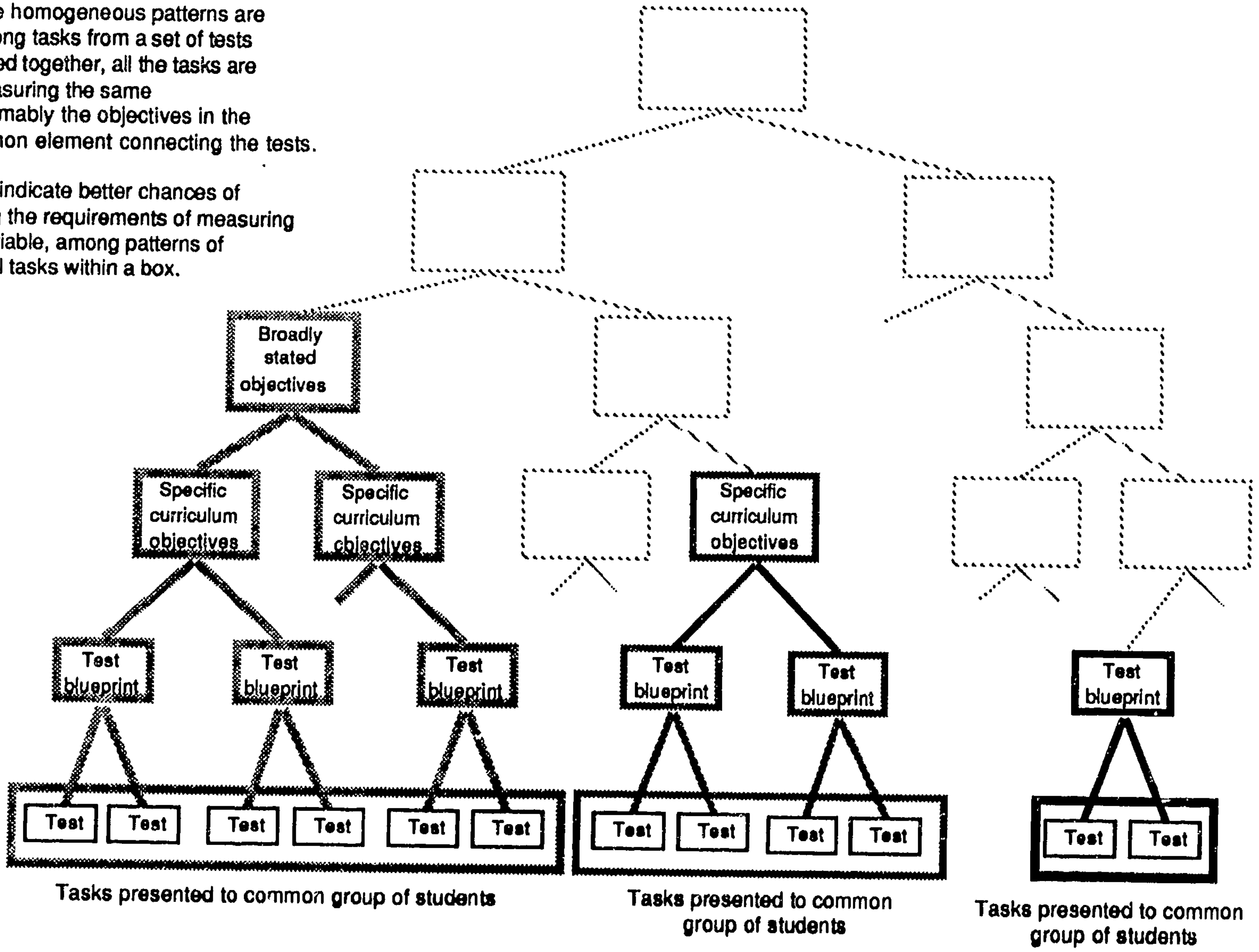
Figure 9 illustrates the linkages that can be forged by this kind of IRT calibration—that is, defining a variable in terms of regularities of behavior patterns across a collection of tasks. The chances that these measurement requirements will be adequately satisfied are better for tests built from the same framework, although it need not happen even here. The required regularities may also be found across sets of tasks written to different specifications, or from different conceptions of competence, but this is even less likely. Whether it happens in a given application is an empirical question, to be answered with the data from a linking study in which students are administered tasks from both assessments.

Figure 9

Calibration in Terms of Behavior

If the requisite homogeneous patterns are observed among tasks from a set of tests being calibrated together, all the tasks are implicitly "measuring the same thing"—presumably the objectives in the highest common element connecting the tests.

Darker boxes indicate better chances of approximating the requirements of measuring a common variable, among patterns of response to all tasks within a box.


Broadly stated objectives
Specific curriculum objectives
Specific curriculum objectives
Specific curriculum objectives
Test blueprint
Test blueprint
Test blueprint
Test blueprint
Test blueprint
Test blueprint
Test
Test
Test
Test
Test
Test
Test
Test
Test
Test
Test
Test
Tasks presented to common group of students
Tasks presented to common group of students
Tasks presented to common group of students

Stone Lifting as a Measure of Strength

Suppose we have a room full of stones, with weights from five to 500 pounds. To learn about people's strength in a way that corresponds to classical test theory, we would select a random sample of, say, 50 stones, and record how many a person could lift. This would give us (1) an estimate of the proportion of stones in the room each person could lift, if he or she tried them all, and (2) a way to rank people in terms of their strength (a norm-referenced inference). It wouldn't tell us whether a particular person would be likely to lift a 100-pound stone (an example of a criterion-referenced inference). A different sample of 50 stones would give the same kind of information. Because one sample of stones might have more heavy ones, we could equate the number of stones lifted from the two samples, just as we equate scores from alternate forms of the SAT.

Comparing peoples' strength in this manner would require strong people to lift very light stones, and the less strong to try heavy ones—both expending their time and energy without telling us much about their strength. A better system would note exactly which stones a person attempts. People can generally lift stones up through a certain weight, have mixed success in a narrow range, then rarely lift much heavier ones. We might characterize their strength by the point at which they have 50-50 chances of lifting a stone. Our accuracy for a particular person would depend on how many stones are in the right range for him or her. Knowing Alice succeeds with 70 and 90, but not 110 and 130, would yield an estimate for her of 100, plus or minus 10. Having 95, 100, and 105 pound stones would make her "test" more precise. We could then give an estimate within five pounds. These same stones would tell us virtually nothing for measuring the strength of Jonathan, who succeeded with 10 and 15 pounds but not 20 or 25.

If we didn't have a chance to weigh the stones ahead of time, we could observe, for a number of

> people, which stones they could lift and which they couldn't. Believing the stones are lined up in an order that applies in the same way to every person, we could use the relative frequencies with which stones are lifted to discover this order, then use it to measure more people for whom we believe the same ordering also applies.

In analogy to stone lifting, Rasch's IRT model for right/wrong test items proposes the probability that a person will respond correctly to an item is a function of (1) a parameter characterizing the person's proficiency and (2) a parameter characterizing the item's difficulty. If the model holds, long and short tests, hard and easy ones, constructed from the pool of items might be "calibrated" like the stones.[12] The estimates for peoples' locations would be scattered around their "true" locations, with the degree of accuracy depending mainly on the number of items administered in the neighborhoods of their true abilities. Comparing peoples' measures with one another or with group distributions supports norm-referenced inferences about individuals. Comparing their scores with item locations supports criterion-referenced inferences. People's IRT measures are approximately unbiased with long tests.[13]

Rasch's IRT model for right/wrong test items proposes the probability that a person will respond correctly to an item is a function of ... the person's proficiency and ... the item's difficulty.

Linking tests with Rasch's IRT model amounts to estimating the locations of tasks from both tests, based on the responses of a sample of people who have taken at least some of each. *If* the IRT model fits the patterns in data well enough, then estimates of students' proficiency parameters are properly calibrated measures on the same scale. A correspondence table could be drawn up that matches the expected scores on the two tests for various values of "true" proficiency. As we saw in the cricket example, the estimate

[12] The likelihood functions for right and wrong responses to each of the items would be estimated from the data, characterizing the evidence each provided about proficiency on the domain of items. The likelihood function induced by a person's set of responses to several items would be the product of the appropriate responses. The highest or most probable point is his or her "maximum likelihood estimate" test score.

[13] They are "consistent" estimates, meaning that they approach being unbiased as the number of observations increases. In the interest of simplicity, I will use the more familiar term "unbiased" loosely, to encompass "consistent."

of an individual student would have about the same expectation from either test, although measurement error might be small with one test and horrendously large with the other.

As we also saw in the cricket example, however, the distribution of unbiased scores for individuals is generally not an unbiased estimate of the distribution of their true scores. The correspondence table described in the preceding paragraph gives the wrong answers about population characteristics such as spread and proportion of people above selected points on the scale. Statistical machinery is available (again, *if* the IRT model holds) to estimate population characteristics whether the test forms are long or short, hard or easy, but questions about group distributions must be addressed with data from the group as an entity. Mislevy, Beaton, Sheehan, and Kaplan (1992), for example, describe how such procedures are used in NAEP. They show how using a demonstrably good estimate for every student in a group can give a demonstrably bad estimate about the characteristics of the group as a whole.

...the distribution of unbiased scores for individuals is generally not an unbiased estimate of the distribution of their true scores.

...using a demonstrably good estimate for every student in a group can give a demonstrably bad estimate about the characteristics of the group as a whole.

I keep saying "*if* the IRT model holds." Beyond simply asserting that measurability holds, IRT models give a way to gauge their plausibility in terms of the patterns observed in data. Sometimes students who do well on most of the items miss easy ones, and students who don't do well on most answer some hard ones correctly. Suppose Charlie, Gwyn, and Peter all answer three of five items correctly. Peter and Gwyn miss items #4 and #5, which were in fact the hardest ones. Unexpectedly, Charlie misses the easiest ones, items #1 and #2. This is like lifting stones most people find heavy but failing to lift ones most people find light. We distrust the assumption that the items line up for Charlie in the same way they do for Gwyn and Peter and question using his score to compare him to his peers. Does Charlie exhibit an atypical response pattern because he uses a different method to solve the problems than most of the other students? The patterns in the observed behavior of these students with the same score are so discrepant as to defy the claim that they are measuring the same thing for everyone. Some variation is expected under the model, and statistical tests help determine whether an individual's response pattern is surprising in this sense or if a particular item tends to be unexpectedly easy or difficult for particular groups of people.

Since an IRT model is never "the truth" anyway, these concerns must be addressed as matters of degree and practical consequence. If one acted as if the measurement model were true, what errors would result, and how serious would they be, for which kinds of inferences?

We have learned that IRT models tend to fit better if the items are more homogeneous and the people are more homogeneous with respect to what the items require. For items, this homogeneity includes content, the way in which they are presented—item formats, timing conditions, order of the items, and so on—and the uses to which they will be put, which affect student motivation. Estimates of group averages can change more from these causes than from a year's worth of schooling (Beaton & Zwick, 1990)! For students, differences in cultural backgrounds and educational experiences that interact with the item content can make it impossible to define a single common measure from a collection of items—the items tend to line up differently for members of different groups. Merely standardizing the conditions of observation is not sufficient for test scores to support measures. Exactly the same items, settings, and timings can lead to different characteristic patterns among students with different backgrounds, muddling the meaning of test scores and subverting comparisons among students.

...IRT models tend to fit better if the items are more homogeneous and the people are more homogeneous with respect to what the items require.

Exactly the same items, settings, and timings can lead to different characteristic patterns among students with different backgrounds, muddling the meaning of test scores and subverting comparisons among students.

Suppose, for example, the tasks in Test X are built around the skills Mathematics Program X emphasizes, while the tasks in Test Y are built around the skills Mathematics Program Y emphasizes. A combined test with both kinds of items might approximately define a variable among Program X students only or among Program Y students only. The combined test used with all students together would produce *scores*, and we could even construct parallel test forms and successfully equate them. But because of the *qualitative* difference between students' profiles of performance for X and Y items, the scores could *not* be thought of as measures on a single well-defined variable. We can attack this problem in two ways:

- Perhaps subsets of the tasks from the two tests cohere, so that one or more portions of different assessments can be calibrated to variables that extend over groups.

- If an IRT model holds across all tasks within each student group separately, but not across groups, we might think of the tests as measures of two distinct variables rather than two alternative measures of the same variable. Rather than calibrating two tests to the same single variable, we would study the relationships among two (or more) distinct variables. This is the topic of the section on "projection."

Calibration in Educational Assessment, Case 3: Item Response Theory for Judgments on an Abstractly Defined Proficiency

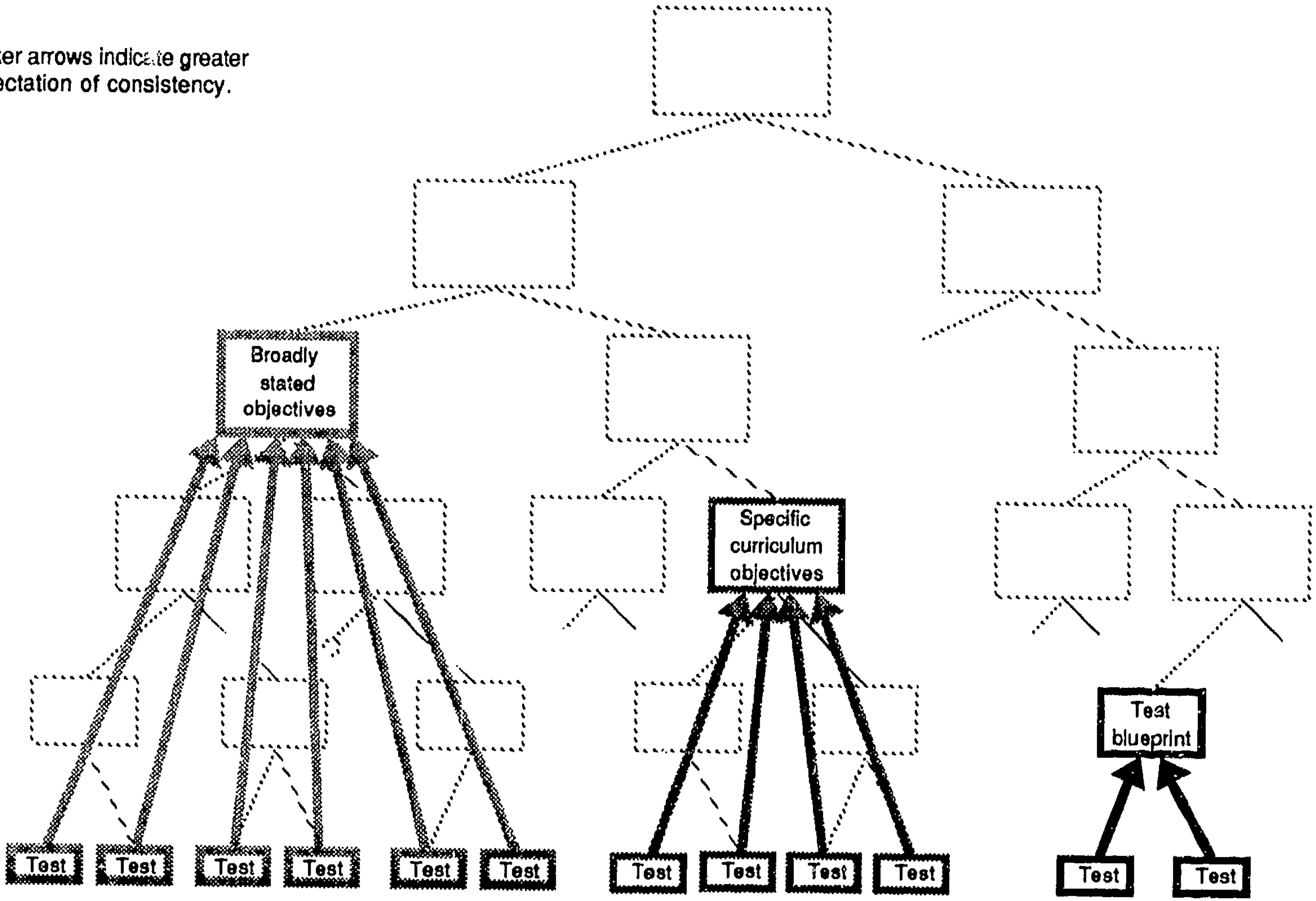
In recent years, IRT modeling has been extended from right/wrong items to ratings and partial-credit data (see, e.g., Thissen & Steinberg, 1986, for a taxonomy of models). An opportunity exists to map judges' ratings of performance on complex tasks onto a common scale, from which properly calibrated measures of student performance might be derived. As with IRT for right/wrong items, an opportunity also exists to discover that performances over a given collection of tasks cannot be coherently summarized as values on a single variable! This process can be attempted at different levels on the hierarchy of competence definitions, as suggested by Figure 10. For reasons discussed below, linking in this manner is more likely to succeed for tasks sharing a more focused conception of competence, that is, more likely to succeed with the grouping of tests at the right of Figure 10 than with those at the left.

The ACTFL language proficiency guidelines can be used to illustrate those points. The calibration process might be carried out within a single language or for performances across several languages, the latter representing a greater challenge. One or more judges, observing one or more interviews with a student, would rate students' accomplishments in terms of ACTFL levels. Students would be characterized in terms of their tendencies to perform at levels of the guidelines, a variable at a higher level of abstraction than the sample of their actual performances. Raters would be characterized in terms of their harshness or leniency. Interview topics or settings, possibly tailored and certainly interpreted individually for individual students, would be characterized in terms of their difficulty.

Compared to multiple-choice or true/false items, a judge rating performances in terms of an abstract definition of competence faces a range of student responses. Mapping what he or she sees to one or more summary ratings is no easy task. Consider putting the ACTFL language proficiency guidelines into operation. The generic descriptions are meant to signify mileposts along the path to competence in learning

Figure 10

Calibration in Terms of Judgment at a Higher Level of Abstraction



Darker arrows indicate greater expectation of consistency.

a foreign language. As they stand, however, the guidelines are just so many words, open to interpretation when applied to any particular performance. They acquire meaning only through examples and discussion: examples for the generic guidelines from different languages, to facilitate meaningful discussion across languages; examples and more specific guidelines within each language, to promote common interpretations for performance among students who will be compared most directly; and examples to work through, to help raters "make the guidelines their own" in ACTFL's intensive four-day training sessions.

...hard work and attention to unusual ratings can bring a judging system to what Deming calls "statistical control."

Even after successful training, judges inevitably exhibit some variations in characterizing any given performance. In a specific setting, however, hard work and attention to unusual ratings can bring a judging system to what Deming calls "statistical control": The extent of variation among and within judges lies within steady, predictable ranges. The more narrowly defined a performance task is, the easier it will usually be for judges to come to agreement about the meanings of ratings, through words or examples, and the more closely their judgments will agree. The Motion Picture Academy might have less trouble selecting the best actress of the year, for example, if all actresses had to play Lady Macbeth in a given year. Of course, making judging easier in this way destroys the intention of recognizing unique, individualistic, and often unanticipated kinds of excellence. So instead, the Academy appropriately tackles the tougher task of judging very different actresses, who take on different kinds of challenges over a broad range of roles.[14]

The Motion Picture Academy might have less trouble selecting the best actress of the year, for example, if all actresses had to play Lady Macbeth in a given year. Of course, making judging easier in this way destroys the intention of recognizing unique, individualistic, and often unanticipated kinds of excellence.

We can expect the typical amount of interjudge variation to increase as student competence is construed more abstractly or as the range of ways it might be manifest broadens. This variation leads to larger measures of uncertainty for students' scores (i.e., ratings induce more dispersed likelihood functions). A model such as Linacre's extension of IRT (1989) to individual raters and other facets of the judging situation helps identify unusual ratings—an interaction between a judge and a performance more out of synch with other ratings of that performance and other ratings by that

[14] It would be fascinating to analyze Academy members' ratings and explore their rationales in detail. To what degree do we find systematic difference related to "schools of acting?" Do voters' approaches and actors approaches interact? Does the plot or message of a movie influence voters?

judge than would be expected, given the usual distribution of variation among raters and performances. Relaying this information back to judges helps them come to agreement about criteria and helps ensure quality control for high-stakes applications.

Experience suggests that, if all other things are equal, the greater the degree of judgment demanded of raters, the more uncertainty is associated with students' scores. The contrast with multiple-choice items can be dramatic. This latitude may be exactly what we want for instructing individual students, because it expands the richness, the variety, the individualization—the *usefulness*—of the assessment experience for the teacher and the student. High-stakes applications in which a common framework of meaning is demanded over many students and across time may prompt us to develop more constrained guidelines; to break scoring into multiple, more narrowly defined rating variables; to train raters more uniformly; or to increase the number of raters per performance, the number of performances per student, or both.

...the greater the degree of judgment demanded of raters, the more uncertainty is associated with students' scores.

This latitude may be exactly what we want for instructing individual students, because it expands the richness, the variety, the individualization—*the usefulness*—of the assessment experience for the teacher and the student.

Comments on Calibration

Educational assessments can be linked through calibration if the evidence each conveys can be expressed in terms of a likelihood function on a common underlying variable. The expected score of a student will be the same on any of the assessments he or she can appropriately be administered, although there may be more or less evidence from different assessments or for different students on the same assessment. One route to producing assessments that can be calibrated is to write them to blueprints that are the same except for having more or fewer tasks. Alternatively, responses to a collection of multiple-choice or judgmentally scored tasks may satisfactorily approximate the regularities of an item response theory model. Arrangements of these tasks in different tests for different students or different purposes can then be calibrated in terms of a common underlying variable.

Educational assessments can be linked through calibration if the evidence each conveys can be expressed in terms of a likelihood function on a common underlying variable.

The proviso for the IRT route is that the patterns in data—interactions between students, tasks, and, if they are involved, judges—must exhibit the regularities of the measurement model. Irregularities such as when, compared to other tasks, some tasks are hard for some kinds of students but relatively easy for other kinds, are more likely to arise when we analyze more heterogeneous collections of tasks and students. IRT statistical machinery can be used to discover unwanted task-by-background interactions, explore the degree to which they affect inferences, and help determine whether breaking assessments into narrower sets of tasks fits better with a measurement paradigm. If the full range of assessment tasks don't, as a whole, conform to a measurement model, perhaps smaller, more homogeneous, groupings of tasks will.

Even when assessments do "measure the same thing" closely enough for our purposes, measuring it with different accuracy introduces complications for inferences at the group level. Although we get unbiased answers to questions about individual students and group means with properly calibrated scores, these same scores give biased estimates of population characteristics, such as the spread and the proportion of students above a particular point on a scale. Statistical methods can provide the correct answers to the latter questions, but they are unfamiliar, complex, and address the configuration of the group's data as a whole rather than as a collection of individual scores for each student (Mislevy, Beaton, Kaplan, & Sheehan, 1992).

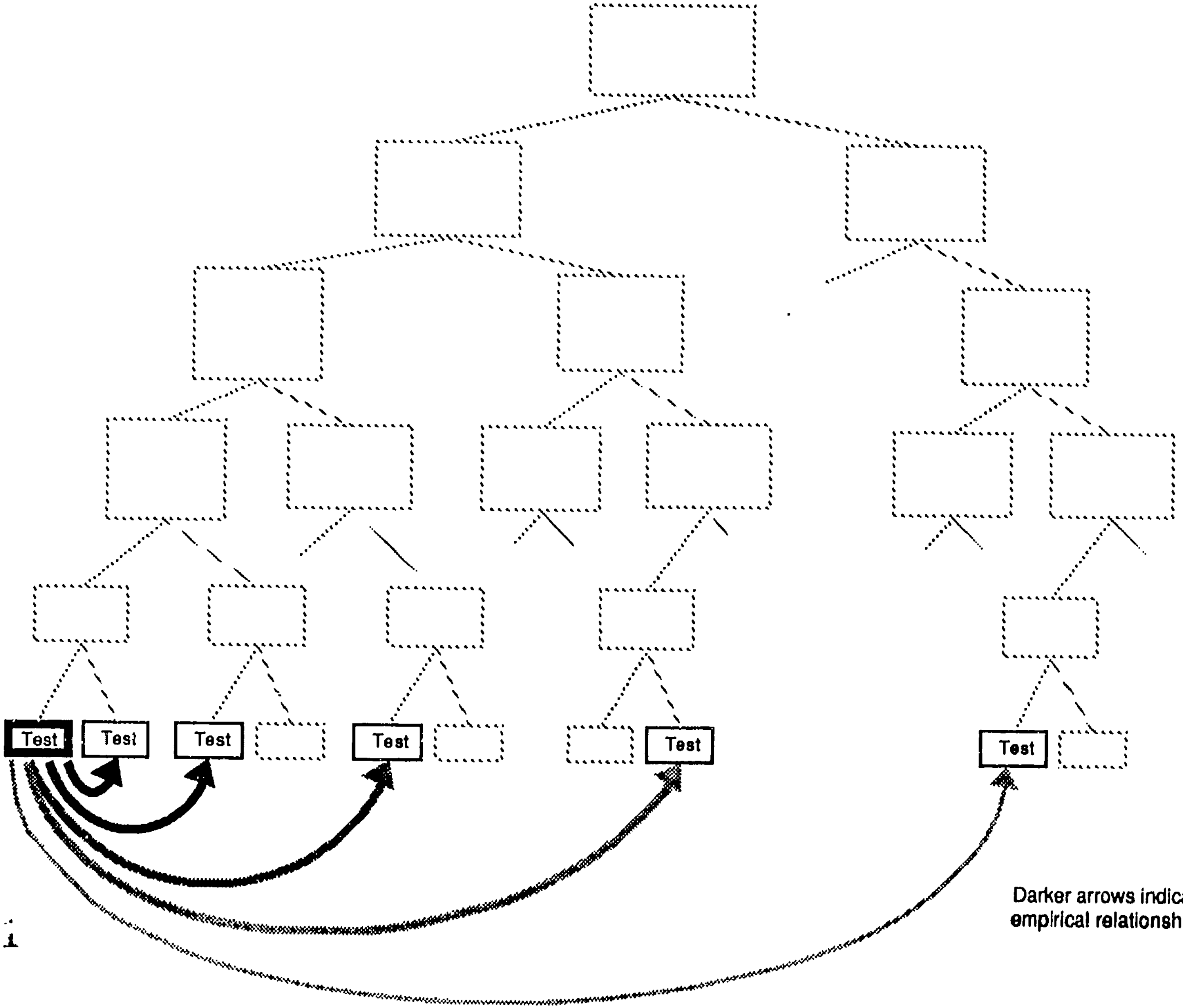
Suppose we have determined that two tests "measure different things." We can neither equate them nor calibrate them...but we may be able to gather data about the joint distribution of scores among relevant groups of students.

Projection

Suppose we have determined that two tests "measure different things." We can neither equate them nor calibrate them to a common frame of reference, but we may be able to gather data about the joint distribution of scores among relevant groups of students. The linking data might be either the performances of a common group of students that has been administered both assessments, or ratings from a common group of judges on performances from both assessments. The links that might be forged are symbolized in Figure 11. We could then make *projections* about how a student with results from Assessment Y might perform on Assessment X, in terms of probabilities of the various possibilities.[15] Projection uses data from a linking study to model the relationship among scores on the two assessments *and other characteristics of students that will be involved in inferences*. I highlight this phrase because "measuring the same thing" means we don't have to worry about these other characteristics, but when "measuring different things," we do. The relationships among scores may change dramatically for, say, students in different instructional programs. As the following examples will show, ignoring these differences can distort inferences and corrupt decisions.

Figure 11

Empirical Linkages for Projection



Darker arrows indicate stronger empirical relationships are expected.

If Ming Mei takes Test Y, what does this tell us about what her score might have been on Test X? While the statistical machinery required to carry it out and gauge its accuracy can be complex, the basic idea behind projection is straightforward: Administer both tests to a large sample of students "suitably similar" to Ming Mei. The distribution of X scores of the people with the same Y score as Ming Mei is a reasonable representation of how we think she might have done. The same idea applies when both X and Y yield multiple scores or ratings. The phrase "suitably similar" is the joker in this deck.

Projection uses data from a linking study to model the relationship among scores on the two assessments and other characteristics of students that will be involved in inferences.

Projection in Physical Measurement

When two instruments measure different qualities, we can study the relationships between their values among a collection of objects and see how these relationships can depend on other properties of the objects. This provides a framework for evaluating information from one variable about conjectures phrased in terms of a different variable. Given a new group of objects with measures on only one variable, we can use the results of the investigation to express our expectations about their values on the second variable.

...the basic idea behind projection is straightforward: Administer both tests to a large sample of students "suitably similar" to Ming Mei.

Linking Temperature and Latitude

Cities closer to the equator tend to be warmer than cities farther away from it. This pattern shows up in our sample of cities in the U.S., as seen in Figure 12. The vertical axis is true temperature. The horizontal axis is negative latitude; left means more northern and right means more southern. The correlation between temperature and southernness is .62, not impressive for indicators of temperature but a fairly high figure for educational test scores—a bit higher, for example, than the correlation between multiple-choice and free-response sections of typical Advanced Placement examinations.

What would we anticipate about the temperature of a city if we knew its latitude? We'd look straight up the graph from that point. For cities around 40° latitude,

[15] A best single guess is a prediction. Predictions are sometimes used to link assessments, but predictions are a special result of the analyses required for projection. We therefore discuss the more general case.

for example, temperatures cluster between 72° and 77°, averaging about 75°. There are also four cities at about 37° latitude. Two have temperatures around 80° while two have temperatures around 65°, for an average of 72.5°. A huge spread, in a strange pattern, but it does represent what we know so far.

How would we use this information linking a sample of cities by temperature and latitude to make inferences about the temperatures of a new sample of cities, for which we know only latitude? For a city at any given latitude, the spread of temperatures for cities at this latitude in the linking sample represents our knowledge. This is the conditional distribution of temperature given latitude. It is used differently for inferences about individual cities and about the new sample as a whole:

- For an individual city, we might use the center of the conditional distribution as a single-point prediction, if we have to give one. The curved line in Figure 12 gives a smoothed version of the centers of the conditional distributions. This line would give a point prediction of 74° for a city at 40° latitude, and a prediction of 76.5° for a city at 37° latitude. Because all the predictions would fall along this line, we'd understate the spread we'd expect in a new sample of cities.

- For the sample as a whole, we use the entire predictive distribution of each city in the new sample. That is, for each city with 40° latitude, anticipate a potential spread between 72° and 77°, centered around 75°. Doing this for every city in the new sample builds up our best guess as to the distribution of temperatures in the sample, taking into account both the features and the limitations of our knowledge.

Let's look more closely at the relationship between temperature and latitude by seeing just which cities are where in the plot. Figure 13 is just like Figure 12, except that three regions of the country have been distinguished: the west coast, the southeast, and the rest (northeast and central). We see that there is very little relationship between latitude and temperature within the west coast cities (Pacific

Figure 12
True Temperatures Plotted Against Latitude

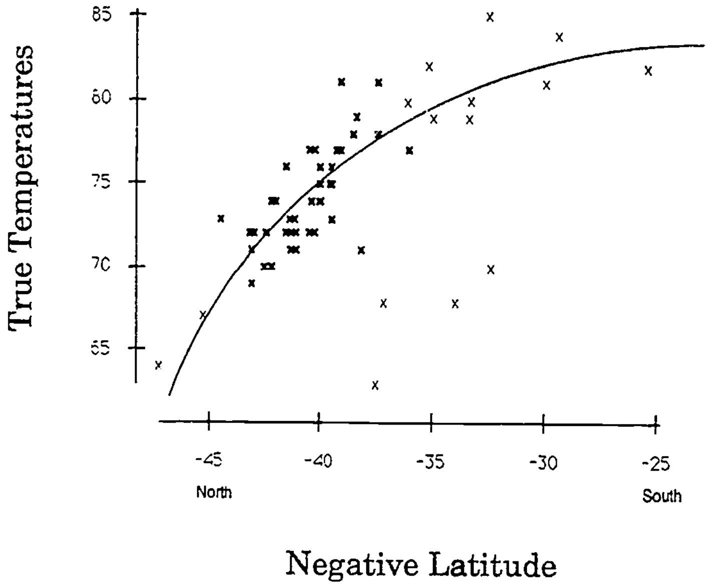
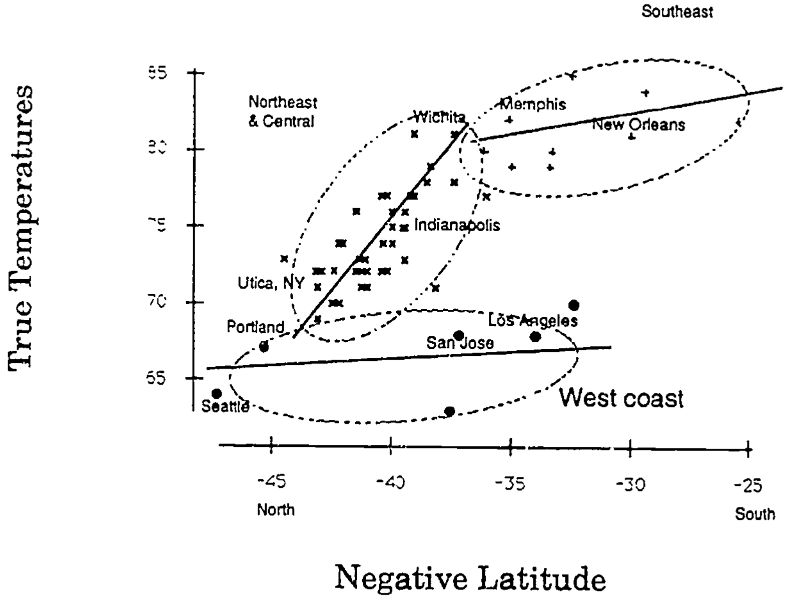


Figure 13
True Temperatures Plotted Against Latitude, with Regions Distinguished



breezes keep them *all* cool!) or within the southeastern cities (they are *all* warm!). The strength of the relationship came mainly from the central and northeast regions. The relationship is strong within this region; the correlation is up to .70, even though there is less variation in latitude. Our projections are much different if we take region into account. Consider the following:

- What temperature would we expect for a city at 40° latitude? If it is in the northeast or the central states, our previous best prediction of 75° is still pretty good—mainly because, we now see, that the only cities in the original sample at 40° latitude were in this region. But if the new city is on the west coast, 75° is *not* a good estimate. A *better* estimate is 67°, the average of all the west coast cities, *no matter how far north or south they are*. And asking what its temperature would be if it were a southeastern city with a latitude of 40° doesn't make sense; by definition, a city that far north can't even be in the southeast!

- What would we expect about a city at 37° latitude? If it's in a northeast or central state, probably about 80°, like Wichita and Richmond; if it's on the west coast, again it's probably about 67°, like all the other west coast cities. The huge spread for cities at this latitude in our original sample was the result mainly of whether or not they were on the west coast.

Projection in Educational Assessment

Two assessments can differ in many ways, including the content of questions, the mode of testing, the conditions of administration, and factors affecting students' motivation. Because each of these changes affects the constellations of skills, strategies, and attitudes each student brings to bear in the assessment context, some students will do better in one than another. In the projection approach to linking assessments, a sample of students is administered all of the assessments of interest (or interlocking portions of them) so we can estimate their joint distribution in that sample. From the resulting estimated distributions, we can answer questions such as, "If I know Ming Mei's results on Assessment Y, what probability distribution represents my expectations for her on Assessment X?" The answer can be applied to an inference we'd have made using Assessment X results, had they been available. The spread of this projected distribution represents the added uncertainty that must be added to the usual measurement uncertainty associated with performance within a given method—using just the best guess would overstate our confidence. We can expect that the more differences there are among assessments' contents, methods, and contexts, the weaker the association among them will be.

In the projection approach to linking assessments, a sample of students is administered all of the assessments of interest...so we can estimate their joint distribution in that sample.

The relationship between assessments can differ systematically for students with different backgrounds, different styles, or different ways of solving problems. Recall the math tests, tailored to different math programs, containing items that could not be calibrated to a single underlying variable. The students instructed from one perspective will tend to do better with tasks written from the same perspective. When we speculate on how well Ming Mei would fare on Assessment X, we look at the X score distribution of "suitably similar" students with the same Assessment Y score(s) as hers.

But suppose the distributions differ for students who have studied in Program X and those who have studied in Program Y. The answer to "What are my expectations for Ming Mei on Assessment X if I know her results on Assessment Y *and that she has studied in Program Y?*" will then differ considerably from the answer to "What are my expectations for Ming Mei on Assessment X if I know her results on Assessment Y *and that she has studied in Program X?*" An answer that doesn't take program study into account averages over the answers that would apply in the two, in proportion to the number of students in the linking sample who have taken them. This gives a rather unsatisfactory statement about the competence of Ming Mei as an individual, for either a high-stakes decision about her accomplishments to date or a low-stakes decision to guide her subsequent instruction. We thus arrive at the first desideratum for linking assessments through projection:

The relationship between assessments can differ systematically for students with different backgrounds, different styles, or different ways of solving problems.

1. *A linking study intended to support inferences based on projection should include relationships among not only assessments, but also other student variables that will be involved in the inference.*

A similar caveat in projection linking is relationships among assessments can change over time. This factor arises when test results are used for accountability purposes, to guide educational policy decisions. The actions that tend to maximally increase scores on one test may not be the same as those that increase scores on a different test, even though the relationship between students' scores on the two tests within any single time point may be highly related.

As an example, students who do well on multiple-choice questions about writing problems also tend to write well, compared with other students. At a given point in time, results from the two tests would rank students similarly and provide similar evidence for norm-referenced inferences. Actually writing essays in class tends to raise the essay-writing skills of all students; even though ordering students according to multiple-choice and essay performances remains similar. Suppose we estimate the joint relationship between essay and multiple-choice measures at a single point in time, but use only multiple-choice scores to track results *over* time. Our one-shot linking study is structurally unable to predict an interaction between assessments over time. No matter whether essay scores would have gone up or down relative to multiple-choice scores, this analysis projects a given trend in multiple-choice scores to the same trend in essay scores. We thus arrive at the second desideratum for linking assessments through projection:

The actions that tend to maximally increase scores on one test may not be the same as those that increase scores on a different test, even though the relationship between students' scores on the two tests within any single time point may be highly related.

2. *If projection is intended to support inferences about change, and changes may differ for the different assessments involved, then linking studies need to be repeated over time to capture the changing nature of the relationships.*

Another Type of Data for Projection Linking

So far we have discussed projection with assessments that can all be reasonably administered to any of the students of interest. Their results can be objectively scored, judgmental, or a mixture. All students in a sample take all assessments under the simplest design for a linking study, and relationships are estimated directly. Under more efficient designs, students take only selected assessments, or even just portions of them, in interlocking patterns that allow the relationships to be estimated.

A variation for judgmental ratings skips a link in the usual chain of inference. Suppose two frames of reference exist for rating performances, perhaps from different conceptions of competence or different operationalizations of what to look for. If evidence for either would be gleaned from similar assessment contexts, it may be reasonable to solicit judgments from both perspectives on the same set of performances. Given results under one approach to scoring, we might then be able to project the distribution of scores under the other.

Under more efficient designs, students take only selected assessments, or even just portions of them, in interlocking patterns that allow the relationships to be estimated.

As an example, Table 6 (based on Linn, n.d.) shows the joint distribution of scores assigned to essays Oregon students wrote for their assessment, as evaluated by the standards of the Oregon "Ideas Score" rubric and Arizona's "Content Score" rubric. Each row in this table roughly indicates the evidence for competence on an aspect of the Arizona framework conveyed by a score on an aspect of the Oregon framework. A more thorough study would be characterized by a larger sample of papers; examination of relationships for interesting subgroups of students, topics, and raters; and, to enable these more ambitious analyses, modeling dominant trends and quantifying sources of variance. Note that the six papers receiving the same Oregon score of 5 received Arizona scores ranging from the lowest to the next to highest—uncertainty that must be taken into account when projecting distributions of Arizona scores from Oregon papers. This amount of uncertainty would give one pause before using Oregon performance for a high-stakes decision for individual students under Arizona standards.

Projection addresses assessments constructed around different conceptions of students' competence, or around the same conceptions but with tasks that differ in format or content.

Comments on Projection

Projection addresses assessments constructed around different conceptions of students' competence, or around the same conceptions but with tasks that differ in format or content. These assessments provide qualitatively different evidence for various conjectures about the competence of groups or individuals. With considerable care, it is possible to estimate the joint relationships among the assessment scores

Table 6
Cross-Tabulation of Oregon Ideas Scores and Arizona Content Scores for the Oregon Middle-School Papers*

	Arizona Content Scores				
Oregon Ideas Scores	0	1	2	3	4
10				3	2
9				3	1
8			2	2	
7			2	4	
6			2	3	
5	1	1	3	1	
4		2	1		
3		3			
2		1			

*Based on Table 17 of Linn (n.d.)

and other variables of interest in a linking study. The results can be used make projections about performance on one assessment from observed performance on the other, in terms of a distribution of possible outcomes. Although it is possible to get a point prediction, no simple one-to-one correspondence table captures the full import of the link for two reasons: (1) using only the best estimate neglects the uncertainty associated with the projection, and therefore with inferences about individual students, and (2) the relationships, and inferences they imply, can vary substantially with students' education and background and can change with the passage of time and instructional interventions.

Projection sounds rather precarious, and it is. The more assessments arouse different aspects of students' knowledge, skills, and attitudes, the wider the door opens for students to perform differently in different settings.

Projection sounds rather precarious, and it is. The more assessments arouse different aspects of students' knowledge, skills, and attitudes, the wider the door opens for students to perform differently in different settings. Moderate associations among assessments can support inferences about how *groups* of students might fare under different alternatives. But projections for *individual* students in high-stakes applications would demand not only strong empirical relationships but vigorous efforts to identify groups (through background investigations) and individuals (through additional sources of information) for whom the usual relationships fail to hold.

Moderation

"Moderation" is a relatively new term in educational testing, although the problem it addresses is not. The goal is to match up scores from different tests that admittedly do not measure the same thing. It differs from projection in the following sense: While projection attempts to characterize the evidence a performance on one assessment conveys about likely performance on another, moderation simply asks for a one-to-one matchup among assessments as to their worth. Moderation is thus an evaluation of value, as opposed to an examination of evidence. We now consider two classes of linking procedures that have evolved to meet this desire. The methods of *statistical moderation* have been developing for more than 50 years as a subclass of scaling procedures (see, for example, Angoff, 1984, and Keeves, 1988). They are statistical in appearance, using estimated score distributions to determine score levels that are deemed comparable. *Social moderation,* a more recent development, relies upon direct judgments.

While projection attempts to characterize the evidence a performance on one assessment conveys about likely performance on another, moderation simply asks for a one-to-one matchup among assessments as to their worth.

Statistical Moderation

Statistical moderation aligns score distributions in essentially the same manner as equating, but with tests that admittedly do not measure the same thing. If two tests can sensibly be administered to the same students, statistical moderation simply applies the *formulas* of equating, without claiming a measurement-theory *justification.* The arrows for projection linkages in Figure 11 illustrate places where statistical moderation might be carried out—except for the two tests on the left that were in fact built to the same blueprint. Applying equating formulas really does equate these tests!

If two tests can sensibly be administered to the same students, statistical moderation simply applies the *formulas* of equating, without claiming a measurement-theory *justification*.

The scales for the SAT Verbal and Mathematics scores (SAT-V and SAT-M) illustrate a simple example of statistical moderation. In April 1941, 10,654 students took the SAT. Their SAT-V and SAT-M formula scores (number correct, minus a fraction of the number wrong) were both transformed to an average of 500 and a standard deviation of 100. Tables of correspondence thus mapped both SAT-M and SAT-V formula scores into the same 200-800 range. An SAT-V scale score and an SAT-M score with the same numerical value are comparable in this normed-referenced sense: In the 1941 sample, the proportion of examinees with SAT-V scores at or above this level, and the proportion with SAT-M scores at or above this level, were the same. (Whether similar proportions of this year's sample have these scores is quite a different question, and in general they don't.)

A more complex example involves "moderator tests." Moderator tests are a device for linking disparate "special" assessments taken by students in different programs or jurisdictions, or for different reasons—for example, German tests for students who study German, and physics tests for students who study physics. Two scores are obtained from each student in a linking study: one for the appropriate special assessment, and one on a "moderator test" that all students take. Scores on the moderator test are used to match up performance on the special tests. The rationale is articulated in *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests* (Donlon & Livingston, 1984, pp. 21) :

> *The College Board offers 14 different achievement tests.... If the scores are to be used for selection purposes, comparing students who take tests in different subjects, the score scales should be as comparable as possible. For example, the level of achievement in American history indicated by a score of 560 should be as similar as possible to the level of achievement in biology indicated by a score of 560 on the biology test. But what does it mean to say that one student's achievement in American history is comparable to another student's achievement in biology? The Admission Testing Program's answer to this question, which forms the basis for scaling the achievement tests, is as follows. Suppose student A's relative standing in a group of American history students is the same as student B's relative standing in a group of biology students. Now suppose the group of American history students is equal to the group of biology students in general academic ability. Then it is meaningful to say that student A's achievement in American history is comparable to student B's achievement in biology.*
>
> *The groups of students who choose to take the different Achievement Tests, however, cannot be assumed to be equal in general academic ability. Their SAT scores often provide evidence that they are not.. . Obviously, the differences are quite large in some cases and cannot be disregarded.*

Donlon and Livingston go on to describe the procedures used to accomplish the scaling.[16] An extreme sports example illustrates the basic idea, using the simpler scheme Keeves (1988) describes for linking grades from different subjects:

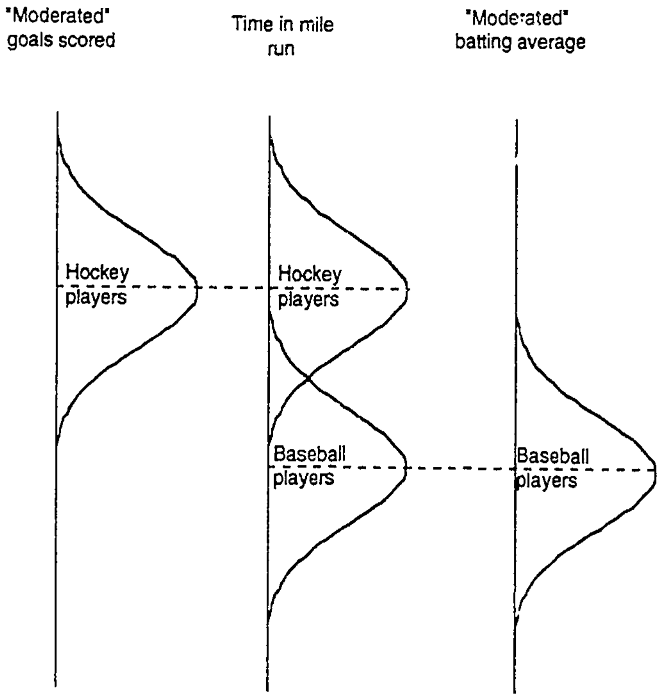
Using Mile-Run Times to Moderate Batting Averages and Goals Scored

> To moderate baseball batting averages and hockey goals scored using mile-run times, we would first obtain batting averages and run times from a group of baseball players, and goals scored and run times from a group of hockey players. Within the baseball players, the best batting average is matched to the best run times, the average to the average, and so on; a similar procedure is used for goals scored and run times within hockey players. To find out what number of goals scored is comparable to a batting average of .250, we (1) find out what proportion of the baseball players in the baseline group hit below this average (say its 60%); (2) look up the mile-run time that 60% of the baseball players were slower than (say its 5:10); (3) look up how many of the hockey players were slower than 5:10 in the mile (say its 30%); and (4) look up how many goals scored that 30% of the hockey players were below (say its 22).
>
> Suppose hockey players tend to be faster runners than baseball players. Figure 14 shows the linkages that result from goals scored to run-

[16] First, relationships among the SAT-V, SAT-M, and an Achievement Test are estimated from an actual baseline sample of students. Then, projection procedures are used to predict the distribution of a hypothetical "reference population" of students who are all "prepared" to take the special area test (i.e., have studied biology, if we are working with the biology test) and have a mean of 500, a standard deviation of 100, and a correlation of .60 on the regular SAT sections. That is, the same relationship among the SAT tests and the Achievement Test observed in the real sample is assumed for the hypothetical sample, which could have a mean higher or lower thar 500 and a standard deviation higher or lower than 100. The projected special-test raw-score distribution of the hypothetical group is transformed to have a mean of 500 and standard deviation of 100.

ning times, and running times to batting averages. This procedure would be consistent with the arguable premise that running times measure athletic ability, and hockey players, because they run fast, would have high batting averages if they were baseball players. The judgment of the relative value of batting and goal-scoring skills is implicit in the choice of the moderator test.

Figure 14
Moderated Distributions of Goals Scored and Batting Averages



Variations in linking samples and moderating tests have no effect on norm-referenced comparisons of performances *within* special area tests, but they can affect norm-referenced comparisons from one special area to another markedly. Carrying out the procedures of statistical moderation with different samples of students, at different points in time or with different moderator tests, can produce markedly different numerical links among tests. Particular choices for these variables can be specified as an operational definition of comparability, but moderated

We would have little confidence in a comparison of, say, subgroup means across "moderated" test scores unless it held up under a broad range of choices for linking samples and moderator tests.

scores in and of themselves offer no clue as to how much the results would differ if the choices were altered. The more the tests vary in content, format, or context, the more the results will vary under alternative moderation schemes. A sensitivity study compares results obtained under different alternatives, revealing which inferences are sensitive to the choices. We would have little confidence in a comparison of, say, subgroup means across "moderated" test scores unless it held up under a broad range of choices for linking samples and moderator tests.

In an application that uses a moderating test, the test's specifications determine the locus of value for "comparable worth." In our sports example, baseball players' performances were assigned lower values than hockey players' goals scored, simply because hockey players ran faster. In a serious educational context, consider the use of statistical moderation to link disparate educational assessments from clusters of schools through NAEP. To the degree that focusing on skills not emphasized in NAEP trades off against skills that are, this arrangement would work in favor of clusters whose tests were most closely aligned with NAEP, and against clusters whose content and methodology departed from it.

A more subtle variation of statistical moderation is applying an IRT model across collections of tasks and students in the face of the unwanted interactions we discussed in the section on calibration. If a composite test consisting of tasks keyed to Program X and Program Y is calibrated with responses from students from both programs, task-type may interact with student-program. Consequential interactions launch us into the realm of statistical moderation. Comparisons among students and between groups could turn out differently with a different balance of task types or a different balance of students in the calibration group. If the interaction is inconsequential, however, inferences could proceed under the conceit of a common measurement model; few inferences would differ if the balance of items or the composition of the calibration sample were to shift. Determining the sensitivity of a link to these factors is a hallmark of a responsible IRT linking study.

Social Moderation

Social moderation calls for direct judgments about the comparability of performance levels on different assessments. This process could apply to performances in assessment contexts that already require judgment or score levels in objectively scored tests. As an example, Wilson (1992) describes the "verification" process in Victoria, Australia. Samples of students' performances on different assessments in different localities were brought together at a single site to adjudicate "comparable" levels of performance.

Social moderation calls for direct judgments about the comparability of performance levels on different assessments. This process could apply to performances in assessment contexts that already require judgment or score levels in objectively scored tests.

Auxiliary information on common assessment tasks can be solicited to supplement direct judgment. For example, assessments of two school districts might contain tasks unique to each, but a common core may be present in both. If so, it would then be possible to compare the score distributions on unique tasks of students from both districts who had the same levels of performance on the common tasks. If nothing more were done but to simply align these distributions, we would have an instance of statistical moderation. If judgments were used to further adjust the resulting matchup on the basis of factors left out of statistical moderation, such as the relevance of the common tasks to the unique tasks, we would have social moderation.

Scoring the Decathlon

The decathlon is a medley of 10 track and field events: 100-meter dash, long jump, shot put, high jump, 400-meter run, 110-meter hurdles, discus throw, pole vault, javelin throw, and 1,500 meter run. Conditions are standardized within events, and it is easy to rank competitors' performances within each. To obtain an overall score, however, requires a common scale of value. This is accomplished by mapping each event's performance (a height, a time, or a distance) onto a 0-1,000 point scale, where sums are accumulated and overall performance is determined. A table was established in 1912 for the decathlon's first appearance in the Olympics by the International Amateur Athletic Federation (IAAF) by a consensus among experts. Performances corresponding to then-current world records were aligned across events, and lesser performances were awarded lower scores in a manner the committee members judged to be comparable.

The IAAF revised the decathlon scoring table in 1936 and 1950 to reflect improvements in world-class performance and to reflect different philosophies of valuation. All of these earlier tables emphasized excellent performance in individual events, so that a superior performance in one event could more than offset relatively poor showings in several others. By scaling down the increases at the highest levels of performance, revisions in 1964 and 1985 favored the athlete who could perform well in many events.

Table 7 shows the performance and total scores of the top eight contenders from the 1932 Olympics in Los Angeles. Their scores under both the 1912 tables, which were in effect at the time, and the 1985 tables are included. The 1985 totals are all lower, reflecting the fact that the top performances in 1932 were not as impressive as those in 1985. James Bausch's performances won the gold medal, but he would have finished behind Akilles Järvinen had the 1985 tables been used. With the 1912 tables, Bausch's outstanding shot put distance more than compensated for his relatively slower running times.

Like statistical moderation, social moderation is founded on a particular definition of comparability, defined in terms of a given process, at a given point in time, with a given set of people.

Like statistical moderation, social moderation is founded on a particular definition of comparability, defined in terms of a given process, at a given point in time, with a given set of people. The set of people now, however, is the sample of "social moderators"— corresponding to the IAAF committees that draw up decathlon tables—rather than a sample of students who take multiple assessments. No built-in mechanism indicates the uncertainties associated with these choices. Like statistical moderation, social moderation can also be supported by sensitivity studies. How much do the alignments change if we carry them out in different parts of the country? How about with teachers versus content-area experts, or students, or community members carrying out the judgments? With different supplemental information to aid the process? Again, we would have little confidence in inferences based on moderated test scores unless they held up under a broad range of reasonable alternatives for carrying out the moderation process.

Table 7
Decathlon Results from the 1932 Olympics*

Names		Throwing Events			Jumping Events			Running Events				Total Points	
		Discus Throw	Shot Put	Javelin Throw	High Jump	Long Jump	Pole Vault	110 Hurdles	100 Meters	400 Meters	500 Meters	1912 Table	1985 Table
1. James Bausch	U.S.	44.58	15.32	61.91	1.70	6.95	4.00	16.2	11.7	54.2	5:17.0	8462	6735
2. Akilles Järvinen	Fin.	36.80	13.11	61.00	1.75	7.00	3.60	15.7	11.1	50.6	4:47.0	8292	6879
3. Wolrad Eberle	Ger.	41.34	13.22	57.49	1.65	6.77	3.50	16.7	11.4	50.8	4:34.4	8031	6661
4. Wilson Charles	U.S.	38.71	12.56	47.72	1.85	7.24	3.40	16.2	11.2	51.2	4:39.8	7985	6716
5. Hans-Heinrich Sievert	Ger.	44.54	14.50	53.91	1.78	6.97	3.20	16.1	11.4	53.6	5:18.0	7941	6515
6. Paavo Yrölä	Fin.	40.77	13.68	56.12	1.75	6.59	3.10	17.0	11.8	52.6	4:37.4	7688	6385
7. Clyde Clifford Coffman	U.S.	34.40	11.86	48.88	1.70	6.77	4.00	17.8	11.3	51.8	4:48.0	7534	6265
8. Robert Tisdall	Irl.	33.31	12.58	45.26	1.65	6.60	3.20	15.5	11.3	49.0	4:34.4	7327	6398

* From Wallechinsky, 1988, The complete book of the Olympics

Comments on Moderation

Moderation should not be viewed as an application of the principles of statistical inference, but as a way to specify "the rules of the game."

Moderation should not be viewed as an application of the principles of statistical inference, but as a way to specify "the rules of the game." It can yield an agreed-upon way of comparing students who differ qualitatively, but it doesn't make information from tests that aren't built to measure the same thing function as if they did. An arbitrarily determined operational definition of comparability must be defended. In *statistical* moderation, this means defending the reasonableness of the linking sample and, if there is one, the moderating test. This definition is bolstered by a sensitivity study showing how inferences differ if these specifications change to other reasonable choices. In *social* moderation, consensual processes for selecting judges and carrying out the alignment constitutes a first, necessary line of defense, which can similarly be bolstered by sensitivity studies.

Conclusion

Can We Measure Progress Toward National Goals if Different Students Take Different Tests?

If we construct assessments carefully and interpret them in light of other evidence, they can provide useful evidence about aspects of students' competencies to guide decisions about instruction and policy.

An educational assessment consists of opportunities for students to use their skills and apply their knowledge in content domains. A particular collection of tasks evokes in each student a unique mix of knowledge, skills, and strategies. Results summarize performances in terms of a simplified reporting scheme, in a framework determined by a conception of important features of students' competence. If we construct assessments carefully and interpret them in light of other evidence, they can provide useful evidence about aspects of students' competencies to guide decisions about instruction and policy.

A given assessment is designed to provide evidence about students for a particular purpose, and in terms of a particular conception of competence. The notion of "linking" Assessment Y to Assessment X stems from the desire to answer questions about students' competence that are cast in the frame of reference that led to Assessment X, when we observe only results from Assessment Y. The nature of the relationships among the methodology, content, and intended purpose of Assessment X and those of Assessment Y determine the statistical machinery necessary to address those questions.

No single score can give a full picture of the range of competencies of different students in different instructional programs.

The notion of monitoring national standards implies gathering evidence about aspects of students' competence within a framework that is meaningful for all students, say in a given grade, in all schools throughout the nation. In any content area, the development of competence has many aspects. No single score can give a full picture of the range of competencies of different students in different instructional programs. Accordingly, multiple sources of evidence—different question types, formats, and contexts—must be considered. Some of these will be broadly meaningful and useful; others will be more idiosyncratic at the level of the state, the school, the classroom, or even the individual student.

No simple statistical machinery can transform the results of two arbitrarily selected assessments so that they provide interchangeable information about all questions about students' competencies.

No simple statistical machinery can transform the results of two arbitrarily selected assessments so that they provide interchangeable information about all questions about students' competencies. Such a strong link is in fact approximated routinely when alternative forms of the SAT or of the Armed Services Vocational Aptitude Battery (ASVAB) are equated—but only because these forms are written to the same tight form and content constraints and administered under standard conditions. The simple and powerful linking achieved in these cases results not from the statistical procedures used to map the correspondence, but from the way the assessments are constructed.

What, then, can we say about prospects for linking disparate assessments in a national system of assessments? First, it *isn't* possible to construct once-and-for-all correspondence tables to "calibrate" whatever assessments might be built in different clusters of schools, districts, or states to provide different kinds of information about students. What *is* possible, with carefully planned continual linking studies and sophisticated statistical methodology, is attaining the less ambitious, but more realistic, goals below:

...it *isn't* possible to construct once-and-for-all correspondence tables to "calibrate" whatever assessments might be built in different clusters of schools, districts, or states to provide different kinds of information about students.

- *Comparing levels of performance across clusters directly in terms of common indicators of performance on a selected sample of consensually defined tasks administered under standard conditions—a "market basket" of valued tasks.* Some aspects of competence and assessment contexts for gathering evidence about them will be considered useful by a wide range of educators, and elements of an assessment system can solicit information in much the same way for all (see sections on equating and calibration, Case 2). A role like this might be appropriate for the National Assessment. Gathering information in this standardized manner, however, can fail to provide sufficient information about aspects of competence holding different salience for different clusters. Common components should be seen as one of many sources of evidence about students' competencies—a limited source for providing evidence about the variety of individual competencies that characterize a nation of students.

- *Estimating levels of performance of groups or individuals within clusters, possibly in quite different ways in different clusters, at the levels of accuracy demanded by purposes within clusters.* Methodologies for accomplishing this under standardized conditions with objectively scored tasks like multiple-choice items have been available for some time. Analogous methodologies for assessment approaches based on judgment are now emerging. Components of cluster assessments could gather evidence for different purposes or from different perspectives of competence, to complement information gathered by common components.

- *Comparing levels of performance across clusters in terms of performance ratings on customized assessments, in terms of a consensually defined, more abstract description of developing competcnce.* When cluster-specific assessment components are designed to provide evidence about the same set of abstractly defined standards (e.g., those of the National Council of Teachers of Mathematics), though possibly from different perspectives or by different methods, it is feasible to map performance in terms of a common construct (see section on calibration, Case 3). Comparisons across clusters are not as accurate as comparisons within clusters based on the same methodology and perspective, nor as accurate as comparisons across clusters on common assessment tasks. The tradeoff, though, is the opportunity to align assessment with instruction without requiring standardized instruction. For the reasons discussed in the section on statistical moderation, it is inadvisable to link unique cluster components soley through their empirical relationships with the common cluster components.

What is possible?

- *Comparing levels of performance across clusters directly in terms of common indicators of performance on a selected sample of consensually defined tasks administered under standard conditions—a "market basket" of valued tasks.*

- *Estimating levels of performance of groups or individuals within clusters, possibly in quite different ways in different clusters, at the levels of accuracy demanded by purposes within clusters.*

- *Comparing levels of performance across clusters in terms of performance ratings on customized assessments, in terms of a consensually defined, more abstract description of developing competence.*

- *Making projections about how students from one cluster might have performed on the assessment of another.*

- *Making projections about how students from one cluster might have performed on the assessment of another.* When students can be administered portions of different clusters' assessments under conditions similar to those in which they are used in practice, we can estimate the joint distribution of results on those assessments. We can then address "what if" questions about what performances of groups or individuals who took one of the assessments might have been on the other (see section on projection). The more the assessments differ in form, content, and context, though, the more uncertainty is associated with these projections, the more they can be expected to vary with students' background and educational characteristics, and the more they can shift over time. Unless very strong relationships are observed and found to hold up over time and across different types of students, high-stakes uses for individuals through this route are perilous.

References

American Council on the Training of Foreign Languages. (1989). *ACTFL proficiency guidelines*. Yonkers, NY: Author.

Angoff, W.H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service, Policy Information Center.

Barton, P.E. (1992). *National standards for education: What they might look like.* Princeton, NJ: Educational Testing Service, Policy Information Center.

Beaton, A.E., & Zwick, R. (1990). *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly.* (No. 17-TR-21). Princeton, NJ: National Assessment of Educational Progress/Educational Testing Service.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Deming. W.E. (1980). *Scientific methods in administration and management.* Course No. 617. Washington, DC: George Washington University.

Donlon, T.F. (Ed.) (1984). *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests.* New York: College Entrance Examination Board.

Donlon, T.F., & Livingston, S. A. (1984). Psychometric methods used in the Admissions Testing Program. In T.F. Donlon (Ed.), *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests.* New York: College Entrance Examination Board.

Edgeworth, F.Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51,* 599-635.

Edgeworth, F.Y. (1892). Correlated averages. *Philosophical Magazine,* 5th Series, *34*, 190-204.

Glaser, R. (1991). Expertise and assessment. In M.C. Wittrock & E.L. Baker (Eds.), *Testing and cognition.* Englewood Cliffs, NJ: Prentice Hall.

Green, B. (1978). In defense of measurement. *American Psychologist, 33*, 664-670.

Gulliksen, H. (1950/1987). *Theory of mental tests.* New York: Wiley. Reprint, Hillsdale, NJ: Erlbaum.

Jöreskog, K.G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models.* Cambridge, MA: Abt Books.

Keeves, J. (1988). Scaling achievement test scores. In T. Husen & T.N. Postlethwaite (Eds.), *International encyclopedia of education.* Oxford: Pergamon Press.

Krathwohl, D.R., & Payne, D.A. (1971). Defining and assessing educational objectives. In R.L. Thorndike (Ed.), *Educational measurement* (2nd Ed.) Washington, DC: American Council on Education.

Lesh, R., Lamon, S.J., Behr, M., & Lester, F. (1992). Future directions for mathematics assessment. In R. Lesh & S.J. Lamon (Eds.), *Assessment of authentic performance in school mathematics.* Washington, DC: American Association for the Advancement of Science.

Linacre, J. M. (1989). *Multi-faceted Rasch measurement.* Chicago, IL: MESA Press.

Linn, R.L. (n.d.). *Cross-state comparability of judgments of student writing: Results from the New Standards Project.* CSE Technical Report 335. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Linn, R.L. (in press). Linking results of distinct assessments. *Applied Measurement in Education.*

Linse, D.J. (1992). System identification for adaptive nonlinear flight control using neural networks. Doctoral dissertation. Princeton University, Department of Mechanical and Aerospace Engineering.

Lord, F.M. (1962). Estimating norms by item sampling. *Educational and Psychological Measurement, 22,* 259-267.

Lunz, M.E., & Stahl, J.A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions, 13,* 425-444.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed.) New York: American Council on Education/Macmillan.

Mislevy, R.J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359-381.

Mislevy, R.J. (1990). Item-by-form variation in the 1984 and 1986 NAEP reading surveys. In A.E. Beaton & R. Zwick, *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly.* Report No. 17-TR-21. Princeton, NJ: Educational Testing Service.

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.

Mislevy, R.J., Beaton, A.E., Kaplan, B., & Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133-161.

Payne, D.A. (1992). *Measuring and evaluating educational outcomes.* New York: Macmillan.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 221-262). New York: American Council on Education/Macmillan.

Porter, A. C. (1991). Assessing national goals: some measurement dilemmas. In T. Wardell (Ed.), *The assessment of national goals. Proceedings of the 1990 ETS Invitational Conference.* Princeton, NJ: Educational Testing Service.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research/Chicago: University of Chicago Press (reprint).

Schum, D.A. (1987). *Evidence and inference for the intelligence analyst.* Lanham, MD: University Press of America.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15*, 201-292.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

Stocking, M.L., Swanson, L., & Pearlman, M. (1991). *Automated item selection (AIS) methods in the ETS testing environment*. Research Report 91-5. Princeton, NJ: Educational Testing Service.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567-577.

U.S. Department of Education. (1990). *National goals for education.* Washington, DC: Author.

Velleman, P.F. (1988). *Data Desk Professional 2.0* [computer program]. Northbrook, IL: Odesta.

Wallechinsky, D. (1988). *The complete book of the Olympics.* New York: Viking Penguin.

Wilson, M.R. (1992). *The integration of school-based assessments into a state-wide assessment system: Historical perspectives and contemporary issues.*.Unpublished manuscript prepared for the California Assessment Program. Berkeley, CA: University of California, Berkeley.