Evaluating the Adequacy of Social-Emotional Measures in Early Childhood

Authors: Rebecca Gokiert[1] (corresponding author,) Rebecca Georgis[2], Melissa Tremblay[3], Vijaya Krishnan[1], Christine Vandenberghe[4], and Clara Lee[2]

[1]Community-University Partnership for the Study of Children, Youth, and Families, Faculty of Extension, University of Alberta, Edmonton, Alberta, Canada

[2]Department of Educational Psychology, University of Alberta, Edmonton, Alberta, Canada

[3]Department of Pediatrics, University of Alberta, Edmonton, Alberta, Canada

[4]Alberta Centre for Child, Family, and Community Research, Calgary, Alberta, Canada

Evaluating the Adequacy of Social-Emotional Measures in Early Childhood

Abstract

Technical adequacy and usability are important considerations in selecting early childhood social-emotional (SE) screening and assessment measures. As identification of difficulties can be tied to programming, intervention, accountability, and funding (e.g., Glascoe, Foster, & Wolraich, 1997) it is imperative that practitioners and decision-makers select appropriate and quality measures from the plethora of measures available. This study systematically reviewed and evaluated the technical adequacy and usability of 10 commonly used SE assessment and screening measures, employing a framework for evaluating selected properties of measures (e.g., reliability, validity). Through this review it was found that there are inadequacies in many commonly used SE measures, deserving the attention of both users and developers.

Key words: Early childhood, Social-Emotional, Assessment measures, Screening measures, Evaluation, Psychometrics

Evaluating the Adequacy of Social-Emotional Measures in Early Childhood

Research across disciplines has confirmed the importance of the early years in setting the foundation for long-term learning, behaviour, and health outcomes (e.g., Shonkoff & Phillips, 2000). One component of healthy early childhood development is social-emotional (SE) competence, defined as cooperative and pro-social behaviour; engagement in positive relationships with parents, peers, siblings, and educators (Raver & Ziegler, 1997); management of aggression and/or conflict; development of self-worth; and ability to regulate emotion and reactivity (Squires, 2003). Difficulties with SE competence in the early years are often associated with later troubles in school readiness and performance, social adjustment, and health (Carter, Briggs-Gowan, & Davis, 2004; McCain, Mustard, & Shanker, 2007).

The early detection, through screening and assessment, of students at-risk for developing SE difficulties can guide early intervention programming and funding, all of which promote healthy development (American Academy of Pediatrics, 2007). Screening is a cost-effective method that can be used by paraprofessionals to identify children who are at-risk of experiencing difficulties in development, consequently requiring additional assessment or intervention (Bricker, Schoen, & Squires, 2004). Assessment is a comprehensive evaluation of a child's functioning that can lead to identification of disabilities or disorders, which typically leads to enhanced educational or clinical services (Henderson & Strain, 2009).

A number of reviews of preschool measures of behaviour and learning have been conducted (e.g., Bracken, Keith, & Walker, 1998; Caselman & Self, 2008; Henderson & Strain, 2009; Humphrey, Kalambouka, Wigelsworth, Lendrum, Deighton, & Wolpert, 2011) all echoing the technical inadequacies of measures within the early years. In a review of 13 third-party instruments for measuring social-emotional functioning, Bracken et al. (1998) concluded that

although none of the instruments met the desired criteria, some such as the *Behaviour Assessment System for Children (BASC)* and the *Vineland* possessed stronger technical qualities than others. Even with higher psychometric standards, more recent reviews have highlighted continued limitations (Bricker et al., 2004; Emmons & Alfonso, 2005).

Professional organizations and researchers (e.g., AERA, APA, & NCME, 1999; Keszei, Novak, & Streiner, 2010; Scientific Advisory Committee of the Medical Outcomes Trust, 2002) have identified a number of key standards that contribute to the technical adequacy of measurement instruments. These include the purpose or conceptual basis of the measure, the standardization process and representativeness of the normative sample, reliability, and validity. A number of researchers have established and applied standards to determine the adequacy of measures (e.g., Bracken et al., 1998; Emmons & Alfonso, 2005; Tyson & Connell, 2009); however, the instruments evaluated and criteria applied have no uniformity, making it impossible to draw any firm conclusions. There is also a need to balance psychometric qualities with usability, the clinical or practical utility of a measure (CanChild Centre for Childhood Disability Research, 2004). Further, the respondent and administrative burden should be examined in light of the time needed to complete the form or interview, reading and comprehension levels for questions, and the level of training or expertise needed to administer the measure and interpret the results (Scientific Advisory Committee on Medical Outcomes Trust, 2002). In light of this, it is important that professionals possess the information that is necessary to critically evaluate all available measures and make informed decisions that are in the best interests of children and their families. The purpose of this study was to review and investigate the technical and usability characteristics of 10 early childhood SE measures and present this information in an easily accessible format.

The 10 SE Measures for Evaluation

The 10 measures were selected from a larger pool of SE measures and are presented in alphabetical order and described in Table 1. The measures selected are commonly used for screening or assessment in early childhood and have a focus on SE competence exclusively. To begin, 78 measures were identified for their suitability in early childhood (birth to 8 years), based on such criteria as availability in English, through an extensive search process that included: (a) electronic databases (e.g., PsycINFO, ERIC, and Medline); (b) test publisher websites; and (c) compendiums on SE measures (e.g., Sosna & Mastergeorge, 2005).

Of the 10 measures selected for evaluation, seven were assessment instruments (ABAS-II, BASC-2, CBCL/1.5-5, ITSEA, SIB-R, Vineland-II, Vineland SEEC) and three were screening tools (ASQ:SE, BITSEA, Greenspan). For the purpose of this review, a*ssessment* is defined as a process of systematic observations and analysis used to deepen understanding of a child's competencies whereas *screening* involves the use of a tool designed to identify children who may be at-risk for social emotional difficulties or developmental delays (American Academy of Pediatrics, 2001; Greenspan & Meisels, 1996; Meisels & Atkins-Burnett, 2000).

[INSERT TABLE 1 HERE]

Criteria for Evaluating the SE measures

The most frequently evaluated technical aspects of measurement tools include, but are not limited to, reliability, validity and standardization. Surprisingly, little to no attention has been paid to the usability considerations of a tool such as cost, readability, test length, completion time, and administration and interpretation training required. For our review, an evaluation

protocol was created, consisting of 16 items across the four domains of reliability, validity, standardization, and usability. Three items from the evaluation protocol were used to evaluate aspects of reliability such as internal consistency, inter-item and test-retest reliabilities, and five to evaluate aspects of validity such as content, criterion, construct, and sensitivity and specificity. The standardization section of the protocol comprised four items evaluating the normative, early childhood, and age-specific samples and stratification (based on such variables as gender and socioeconomic status). Finally, four items were used to evaluate the usability of each measure. Criteria were established for each of the 16 items that comprised the four domains on the protocol (Table 2), and were based on several resources: AERA, APA, & NCME (1999); Bracken et al. (1998); CanChild Centre for Childhood Disability Research (2004); Caselman & Self (2008); Fallon, Westaway, & Moloney, 2008; Reneman, Dijkstra, Geertzen, & Dijkstra, 2009; Sandal, Hemmeter, Smith, & McLean (2005); Scientific Advisory Committee on Medical Outcomes Trust (2002); Sveinbjornsdottir & Thorsteinsson, 2008; and Tyson & Connell (2009).

[INSERT TABLE 2 HERE]

Method

The 10 selected tools were evaluated utilizing the protocol to determine if they met the criterion on a 3-point scale with NE = no evidence, 1= evidence provided but criterion not met, 2 = criterion partially met, and 3 = criterion met. In order to evaluate each measure, we relied primarily on the user manual and the publisher's website, limiting evidence strictly to early childhood (0-5 years). It should be noted that factor analytic studies as a critical source of validity evidence, beyond what was included in the user manual, were not part of this review. To establish our own inter-rater agreement, two researchers reviewed three of the 10 measures (the

mean inter-rater agreement was 84% and rose to 100% after discussion) and a single researcher reviewed the remaining seven measures.

Results

The results are reported separately for each of the domains: reliability, validity, standardization, and usability; and by type of measure: assessment and screening (Tables 4, 5, & 6). Table 3 presents the ratings across the four domains for each of the measures.

[INSERT TABLE 3 HERE]

Reliability

As noted earlier, adequacy of reliability included an evaluation of the measure's internal consistency as well as inter-rater and test-retest reliabilities (Tables 2 & 4).

Assessment measures. All assessment measures met the criterion for internal consistency. All seven showed variability, especially for test-retest and inter-rater. Specifically, Cronbach's alpha was typically reported for internal consistency with some exceptions using the split-half method (e.g., Vineland-II). Pearson product-moment correlation coefficients were used for test-retest and inter-rater reliabilities. The ITSEA was the only measure that met the criterion for inter-rater reliability, the ABAS-II, BASC-2, SIB-R, and Vineland-II only partially met the criterion, and the CBCL/1.5-5 and Vineland SEEC did not meet the criterion. The ABAS-II and the BASC-2 met the criterion for test-retest reliability and the remaining five measures - CBCL, ITSEA, SIB-R, Vineland-II, and Vineland SEEC - only partially met the criterion. Inter-rater reliability coefficients were typically lower (.47-.91), compared to internal consistency (.80-.98) and test-retest reliability (.71-.96).

Screening measures. The ASQ:SE and the Greenspan met the criterion for internal consistency whereas the BITSEA did not. Justification for the lack of internal consistency evidence was noted: "the internal consistency of the Problem and Competence Total scores was not calculated because the items that make up these scales were drawn from many different subscales and were not consistently expected to correlate highly with each other" (Briggs-Gowan & Carter, 2006, p. 33). The BITSEA was the only measure that provided inter-rater reliability evidence, which partially met the criterion. The ASQ:SE and the BITSEA met the criterion for test-retest reliability whereas the Greenspan did not provide any evidence.

[INSERT TABLE 4 HERE]

Validity

As earlier described**,** adequacy of validity evidence included an evaluation of the measure's content, convergent, criterion-related, and construct validity (Tables 2 & 5).

Assessment measures. All measures demonstrated evidence of content validity, which included a combination of: clinical research, reviews of relevant theory and literature, examination of existing measures of similar constructs, expert reviews, field pilot testing, consultations with professionals (e.g., psychologists) and parents, and statistical analysis (e.g., using item response theory models). All measures also provided convergent validity as evidenced by the correlations between the measure and other similar measures, including previous editions; SIB-R met the criterion for convergent validity, the ABAS-II, BASC-2, CBCL/1.5-5, and Vineland-II partially met the criterion suggesting a relatively satisfactory relationship, whereas the ITSEA and Vineland SEEC did not meet the criterion. The ITSEA

presented great variability in terms of the strength of its relationship with other measures, ranging from .34 to .69, whereas the Vineland presented moderate relationships with values fairly close to our criterion. It should be noted, however, that although SIB-R met the criterion, the evidence provided was based primarily on high correlations with its predecessor (SIB) and did not provide any evidence about its relationship with other measures. Overall, convergent evidence was moderate to high (.69-.99) when comparing a measure with its predecessor, and low to high (.18-.90) when comparing a measure with other similar measures. The ABAS-II correlated poorly (.18) with the SIB-R EDF, suggesting divergent rather than convergent evidence.

When evaluating criterion-related validity evidence, all measures provided satisfactory evidence except for the BASC-2 and Vineland-II, which did not provide any evidence for the early childhood age group. Specifically, in the BASC-2 manual the exclusion of clinical-nonclinical preschool group comparisons is justified by stating that, "…groups are not presented for the preschool form because of the imprecision of diagnoses at such a young age" (Reynolds, & Kamphaus, 2004, p.192). Vineland-II provides criterion-related validity evidence but not specific to the preschool group, although criterion-related validity analysis included some younger ages (e.g., ages 3-12). Of the five measures that reported criterion-related evidence, the ABAS, CBCL/1.5-5, and ITSEA also reported effect sizes for the clinical and nonclinical group comparisons.

When evaluating construct validity evidence, all but the SIB-R provided factor analysis that examined the internal structure of the measure. The opposite was found for sensitivity/specificity evidence. Only the ITSEA provided such evidence but only for their maladaptive cluster scores and not for their domain/scale scores.

Screening Measures. All three screening measures demonstrated evidence of content validity, which included a combination of: reviews of relevant theory and empirical literature, existing measures, professional experience, expert reviews, and consultations with parents. Also, all three measures provided satisfactory criterion-related validity evidence. However, only the BITSEA provided convergent validity evidence that was low to moderate and did not meet our criterion. None of the measures provided factor analysis evidence. The ASQ:SE and the BITSEA provided sensitivity/specificity analysis that partially met the criterion. Typically, sensitivity values were lower than specificity values suggesting that measures are better at identifying children who are not at-risk than those who are at-risk.

[INSERT TABLE 5 HERE]

Standardization

Standardization included an evaluation of both an overall and age-specific early childhood sample size, and stratification considerations (Tables 2 & 6).

Assessment measures. Assessment measures had adequate standardization quality as the majority provided evidence that met the criterion for all four items pertaining to standardization. In particular, all assessment measures had an adequate early childhood sample size with the exception of the Vineland-II. The ABAS-II, BASC-2, ITSEA, SIB-R, and Vineland SEEC had at least 100 participants in each age group except for the Vineland-II, which had fewer than 100 and the CBCL/1.5-5 that did not provide any age-specific information. All measures had a relatively representative sample closely matching the U.S. Census data on a minimum of three to a maximum of six variables. All measures reported stratification according to ethnicity/race,

parental education/SES, and geographic region. Five measures (i.e., ABAS-II, ITSEA, SIB-R, Vineland-II, and Vineland SEEC) reported gender as an additional stratification variable and two measures (i.e., Vineland SEEC and SIB-R) also included community size. Vineland SEEC also included age as an additional stratification variable. All measures were standardized with the U.S. Census populations. It is noteworthy, however, that even though all but CBCL/1.5-5 reported sample characteristics that relatively matched the general population, census years spanned widely from 1980 to 2002, which brings into question the accuracy and representativeness of the populations, especially when standardization procedures are done on measures such as Vineland SEEC and SIB-R.

Screening measures. With the exception of the Greenspan, screening measures had adequate overall and age-specific sample sizes. All three measures had a relatively representative sample closely matching census data on a minimum of three to a maximum of five variables. All measures reported stratification according to ethnicity/race, parental education/SES, and geographic region. The ASQ: SE also reported gender as an additional stratification variable. All measures were standardized according to the 2000-2002 U.S. populations.

[INSERT TABLE 6 HERE]

Usability

Usability was evaluated in terms of a measure's reading level, administration time, test length, and administration/interpretation training level requirements (Tables 1 & 2).

Assessment measures. The ABAS-II, BASC-2, CBCL/1.5-5, and ITSEA met the reading level criterion, whereas the SIB-R, Vineland-II, and Vineland SEEC partially met the

criterion, while not providing specific reading level information. However, because they all provided the option of having a professional administer the form as a survey or interview (as opposed to a parent completed form), we gave them partial score suggesting that reading level is acceptable if a professional is the administrator. Administration time ranged from 10 to 60 minutes, with the majority requiring an average of 20 minutes or more. Test length for all assessment measures was typically close to or over 100 items (with the exception of the SIB-R EDF which includes only 40 items). In terms of training, the ABAS-II, SIB-R, Vineland-II and Vineland SEEC required Level B training whereas the ITSEA, BASC-2 and CBCL/1.5-5 required Level C.

Screening measures. All screening measures met all of the usability criteria. Specifically, all have appropriate reading level, require less than 15 minutes of administration time and have fewer than 50 items. In regards to administration/interpretation training demands, the BITSEA and Greenspan required Level B whereas ASQ:SE required Level A training.

Discussion and Conclusions

Systematic reviews and evaluations can contribute to the early childhood measurement literature by providing a critical dialogue on the adequacy and appropriateness of measurement tools. Also, they can provide researchers and test users with a framework for evaluating measures. This study evaluated the technical properties and usability of 10 commonly used early childhood screening and assessment measures against established criteria.

Overall, assessment measures demonstrate stronger psychometric properties and rely on more extensive and rigorous research procedures compared to screening measures, the latter faring better with respect to usability. This should be expected, given the differences in terms of their purposes. Each domain of technical adequacy demonstrated its own areas of strength and

weakness. Assessment measures appear to be reliable, with inter-rater reliability tending to be lower when compared to internal consistency and test re-test. It is noteworthy that the majority of assessment measures that reported on inter-rater reliability did not meet our criterion, deserving further attention from test developers. Screening measures showed greater inconsistency, conforming to existing evidence (e.g., Bracken et al., 1998; Emmons & Alfonso, 2005). Emmons and Alfonso (2005), for instance, evaluated five screening batteries and reported primarily high internal consistencies but moderate to high test-retest reliabilities.

The adequacy of validity evidence demonstrated substantial variability for both types of measures. This should be expected, as validity is a matter of degree, often related to the assumption that other comparative measures have adequate validity. There tends to be great variability in regards to convergent evidence, but more consistency in regards to criterion-related validity. However, it would be beneficial if all measures reported on effect sizes for clinical-nonclinical group comparisons as another indicator of quality. A noteworthy trend is that assessment measures are more likely to report extensive factor analytic studies, whereas screening measures are more likely to report sensitivity/specificity, perhaps due to their differences in nature/purpose. Despite the critical need to know how well a tool differentiates children who are at-risk from those who are not, screening measures do not consistently report such information and if they do, the evidence is inadequate when compared to a criterion, especially for specificity. In fact, the technical inadequacy of screening measures has been previously reported as a barrier to successful early identification and intervention (e.g., Bricker et al., 2004), and some have argued for stronger specificity and sensitivity evidence for screening measures (e.g., Emmons & Alfonso, 2005).

Outdated data and variability in census years were the main problems with standardization quality. Outdated data suggests a need for updates from publishers and cautionary use for service providers, especially in light of increased migration trends in countries like Canada and the U.S. in recent decades. Additionally, the reviewed measures have been developed and standardized primarily with American populations and their transferability to other contexts and populations should be done with a critical lens, as the cross-cultural validity of measures may be limited (Gokiert, Georgis, Chow, & Chiu, 2012).

Generally speaking, the usability of measures has received very limited attention. The present study offers a contribution in this area. Ultimately, the choice of a measure by practitioners relies on a balance of psychometric properties and usability and should be done in light of needs and desired goals. That is to say, as each measure presents its own strengths and weaknesses, test users should critically weigh the available psychometric information before making a choice. “The quality of each criterion as it applies to individual measures should be considered in light of how failure to meet the criterion might affect the usefulness of the scale” (Bracken et al, 1998, p.166), and this should guide decision-making around the selection and use of any measure. More importantly, measures provide only one source of information in the process of identification, diagnosis and intervention, and best practices in assessment and screening, necessitates information or data in multiple formats and from different informants across contexts.

The study has at least two limitations. First, although we examined the presence or absence of factor analytic studies as an indicator of construct validity, due to the limited scope of this paper, we did not thoroughly evaluate factor analysis evidence. We believe that the evaluation of factor analysis can be the focus of a paper of its own. Second, the adequacy of the

reviewed measures was based primarily on the information available in test manuals, and there is a possibility for supplementary evidence addressing areas of strength and weakness in the literature. Yet, as the manual is a primary source of information, increased efforts should be made to better integrate newer empirical evidence with the test manual.

In conclusion, our findings suggest that even though tools provide evidence on technical properties, when evidence is evaluated against stringent criteria, tools appear to fall short in many areas. Further, there appears to be great variability in the type and quantity of reported evidence, especially in regards to validity. This finding raises two questions: (a) how can tool users weigh in on different types of technical evidence provided when selecting a tool, and (b) what is the minimum and necessary evidence that should be reported in test manuals, which are the primary sources of evidence for test users. Lastly, although technical adequacy is highly important, greater emphasis should be given to tool usability. Our findings suggest that assessment tools may fall short in terms of usability and tool selection by practitioners may be done at the expense of technical adequacy. In order to encourage and ensure the use of technically adequate tools, usability must receive greater attention in tool construction and reporting.

References

Achenbach, T. M., & Rescorla, L. A. (2001). *Child Behavior Checklist for Ages 1 ½ to 5*: *Manual for the ASEBA Preschool Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

AERA, APA, NCME (1999). *The Standards for Educational and Psychological Testing*. Washington, DC: AERA.

American Academy of Pediatrics (2001). Developmental surveillance and screening of infants and young children. *Pediatrics, 108*(1), 192-196. doi.org/10.1542/peds.108.1.192

Bracken, B. A., Keith, L. K., & Walker, K. C. (1998). Assessment of preschool behavior and social-emotional functioning: A review of thirteen third-party instruments. *Journal of Psychoeducational Assessment, 16,* 153-169. doi.org/10.1177/073428299801600204

Bricker, D., Davis, M. S., & Squires, J. (2004). Mental health screening in young children. *Infants & Young Children, 17*, 129-144. doi.org/10.1097/00001163-200404000-00005

Briggs-Gowan, M. J. & Carter, A. S. (2006). *Brief Infant-Toddler Social and Emotional Assessment* (BITSEA). San Antonio, TX: Harcourt Assessments.

Bruininks, R. H., Woodcock, R. W., Weatherman, R. F., & Hill, B. K. (1996). *Scales of Independent Behavior – Revised* (SIB-R). Rolling Meadows, IL: Riverside.

CanChild Centre for Childhood Disability Research (2004). *Outcome Measures Rating Form Guidelines*. Hamilton, Ca: McMaster University.

Caselman, T. D., & Self, P. A., (2008). Assessment instruments for measuring young children's social-emotional behavioral development. *Children & Schools, 30*(2), 103-115.

Carter, A. S. & Briggs-Gowan, M. J. (2006). *Infant-Toddler Social and Emotional Assessment* (ITSEA). San Antonio, TX: Harcourt Assessments, Inc.

Emmons, M. R., & Alfonso, V. C. (2005). A critical review of the technical characteristics of current preschool screening batteries. *Journal of Psychoeducational Assessment, 23*, 111-127. doi.10.1177/073428290502300201

Fallon, A., Westaway, J., & Moloney, C. (2008). A systematic review of psychometric evidence and expert opinion regarding the assessment of faecal incontinence in older community-dwelling adults. *International Journal of Evidence Based Healthcare, 6,* 225-259. doi:10.1111/j.1479-6988.2007.00088.x

Glascoe, F. P., Foster, M., & Wolraich, M. L. (1997). An economic analysis of developmental detection methods. *Pediatrics, 99*(6), 830-837. doi.org/10.1542/peds.99.6.830

Gokiert, R. J., Georgis, R., Chow, W., Chui, Y. (2012). Early childhood developmental screening: Does culture play a role? *Health Research Transfer in Alberta: Knowledge Translation Casebook, 3, 22-25.*

Greenspan, S. I. (2004). *Greenspan Social-Emotional Growth Chart: A Screening Questionnaire for Infants and Young Children*. San Antonio, TX; Harcourt Assessment.

Greenspan, S. I., & Meisels, S. J. (1996). Toward a new vision for the developmental assessment of infants and young children. In S.J. Meisels & E. Fenichel (Eds.), *New visions for the developmental assessment of infants and young children* (pp. 11-26). Washington, DC: ZERO TO THREE.

Harrison, P. L., & Oakland, T. (2003). *Adaptive Behavior Assessment System* (2nd ed.) San Antonio, TX: The Psychological Corporation.

Henderson, J., & Strain, P. (2009). Screening for social emotional concerns: Considerations in the selection of instruments. *Roadmap to Effective Intervention Practices*. Tampa, Florida: University of South Florida, Technical Assistance Center on Social Emotional Intervention for Young Children.

Humphrey, N., Kalambouka, A., Wigelsworth, M., Lendrum, A., Deighton, J., & Wolpert, M. (2011). Measures of social and emotional skills for children and young people: A systematic review. *Educational and Psychological Measurement, 71*, 617-637. doi.org/10.1177/0013164410382896

Keszei, A. P., Novak, M., & Streiner, D. L. (2010). Introduction to health measurement scales. *Journal of Psychosomatic Research, 68*, 319-323. doi.org/10.1016/j.jpsychores.2010.01.006

Meisels, S. J. & Atkins-Burnett, S. (2000). The elements of early childhood assessment. In J.P. Shonkoff & S. J. Meilsels (Eds). *Handbook of early childhood intervention* (2nd ed.) (pp. 231-257). Cambridge, MA: Cambridge University Press.

Raver, C. C., & Ziegler, E. F. (1997). Social competence: An untapped dimension in evaluating Head Start's success. *Early Childhood Research Quarterly, 12*, 363-385.

Reneman, M. F., Dijkstra, A., Geertzen, J. H. B., & Dijkstra, P. U. (2009). Psychometric properties of chronic pain acceptance questionnaires: a systematic review. *European Journal of Pain.* Doi:10.1016/j.ejpain.2009.08.003

Reynolds, C. R., & Kamphaus, R.W. (2004). *Behaviour Assessment System for Children* (2nd ed.; BASC-2). Circle Pines, MN: American Guidance Service, Inc.

Sandall, S., Hemmeter, M. L., Smith B. J., & McLean, M. E. (Eds.) (2005). *DEC recommended practices: A comprehensive guide for practical application in early intervention/early childhood special education*. Missoula, MT: Division for Early Childhood.

Scientific Advisory Committee on Medical Outcomes Trust (2002). Assessing health status and quality of life instruments: Attributes and review criteria. *Quality of Life Research, 11*, 193-205. doi.org/10.1023/A:1015291021312

Shonkoff, J., & Phillips, D. (2000). *From Neurons to Neighborhoods: The Science of Early Childhood Development.* Washington, DC: National Academy Press.

Sosna, T., & Mastergeorge, A. (2005). *Compendium of Screening Tools for Early Childhood Social-Emotional Development.* Sacramento, CA: California Institute for Mental Health.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1998). *Vineland Social-Emotional Early Childhood Scales (Vineland SEEC).* United States of America: American Guidance Service, Inc.

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland Adaptive Behavior Scales* (2nd ed.) Survey Forms Manual; Survey Interview Form and Parent/Caregiver Rating Form. Minneapolis, MN: Pearson.

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2006). *Vineland Adaptive Behavior Scales* (2nd ed.) Teacher Rating Form Manual. Minneapolis, MN: Pearson.

Squires, J. (2003). *The importance of early identification of social and emotional difficulties in preschool children.* Prepared for the Center for International Rehabilitation.

Squires, J., Bricker, D., & Twombly, E. (2002). *Ages And Stages Questionnaires: Social-Emotional* (ASQ:SE), A parent-completed, child-monitoring system for social-emotional behaviors. Baltimore: Paul H. Brookes Publishing Co.

Sveinbjornsdottir, S., & Thorsteinsson, E. B. (2008). Adolescent coping scales: a critical psychometric review. *Scandinavian Journal of Psychology, 49,* 533-548. Doi:10.1111/j.1467-9450.2008.00669.x

Tyson, S. F., & Connell, L. A. (2009). How to measure balance in clinical practice: A systematic review of the psychometrics and clinical utility of measures of balance activity for neurological conditions. *Clinical Rehabilitation, 23*, 824-840. doi.org/10.1177/0269215509335018

Table 1. *Descriptive Information for the 10 SE Measures*

Measure Name	Age Range	Purpose	Content	Forms	Administrator Training Level	Administration Time (in Minutes)
Adaptive Behavior Assessment System, 2nd ed. (ABAS-II)	Birth-89 years	A	Adaptive Behavior	Parent Teacher	B	15-20
Ages and Stages Questionnaire: Social-Emotional (ASQ :SE)	3-66 months	S	Social Emotional	Parent	A	10-15
Behavior Assessment System for Children, 2nd ed. (BASC-2)	2 -21 years	A	Behavioral Emotional	Parent Teacher	C	10-20
Brief Infant Toddler Social Emotional Assessment (BITSEA)	12-35 months	S	Social Emotional	Parent	B	7-10
Child Behavior Checklist for Ages 1½ to 5 (CBCL/1.5-5)	1½-5 years	A	Behavioral Internalizing Externalizing Problems	Parent Teacher	C	15-60
Greenspan Social Emotional Growth Chart (Greenspan)	0-42 months	S	Social Emotional	Parent	B	10
Infant Toddler Social Emotional Assessment (ITSEA)	12-36 months	A	Social Emotional	Parent	C	25-30
Scales of Independent Behavior, Revised (SIB-R)	Infancy- 80+ years	A	Adaptive Behavior	Parent Professional-Interview Form	Not specified	45-60 (full form) 15-20 (early development form)
Vineland Adaptive Behavior Scales, 2nd ed. (Vineland-II)	Birth-90 years	A	Adaptive Behavior	Parent Teacher Professional-Interview Form	B	20-60
Vineland Social-Emotional Early Childhood Scales (Vineland SEEC)	Birth-5 years	A	Social Emotional	Parent	B	15-25

Note: Measures are presented in alphabetical order; A=assessment and S= screening

Table 2. *Technical Properties and Usability Evaluation Criteria*

Item	Definition and Criterion
Internal consistency (IC)	Extent to which items measure the same trait/domain as shown by the relationship between items. Criterion: Chronbach's α/ Kuder Richardson-20 ≥0.8 or Kappa /Intraclass Correlation Coefficient (ICC) ≥0.7 (Fallon, Westaway, & Moloney, 2008)
Inter-rater reliability (IR)	Level of agreement between measure ratings obtained from different raters. Criterion: Kappa/ICC ≥0.7 or Pearson ≥0.8 (Squires, 2003)
Test-retest reliability (TR)	Level of consistency between repeated administrations of the measure over a time interval. Criterion: Kappa/ICC ≥0.7 (Reneman, Dijkstra, Geertzen, & Dijkstra, 2009) or Pearson ≥0.8 (Sveinbjornadottir & Thorsteinsoon, 2008)
Content validity (CV)	Extent to which the measure's content is comprehensive and based on theoretical and empirical evidence such as expert judgment, literature review etc. (Reneman et al., 2009)
Criterion validity (CV)	Relationship between measure scores and an external criterion such as clinical group membership. Criterion: statistically significant group difference based on ANOVA / Kruskal-Wallis Test
Convergent validity (CV)	Extent of the relationship between the measure and other measures assessing similar constructs. Criterion: correlations ≥0.7 (Sveinbjornadottir & Thorsteinsson, 2008)
Construct validity (CV)	Extent to which a measure's internal structure (grouping of items to subscales/subdomains and subscales/subdomains to domains and composites) is supported. Criterion: Evidence of Factor Analysis
Sensitivity & Specificity (SS)	Test's ability to differentiate between children who are at-risk (true positive) from those who are not at-risk (true negative). Criterion: sensitivity/specificity ≥ .80 (Squires, 2003)
Early Childhood Standardization Sample (ECS)	Early childhood sample used in the standardization process Criterion: Sample size ≥500
Sample size per age group (SG)	Number of participants in the various 0-5 age groups. Criterion: ≥ 100 subjects per age group
Stratification (St)	Variables used for stratification (i.e., SES, gender, geographical region, ethnicity/race, education). Criterion: ≥3 stratification variables
Reading level (RL)	Reading level of the measure. Criterion: if parent completed, a grade 6 to 8 reading level is desirable; if administered by a professional, a higher reading level is acceptable.
Administration time (AT)	Time needed for completing/administering the measure. Criterion: ≤ 15 minutes
Test length (TL)	Number of measure items. Criterion: ≤ 50 items (Sveinbjornsdottir & Thorsteinsson, 2008)
Administrator's training level (AL)	Training required for administering, scoring, and interpreting the measure. Criterion: Level A or Level B. Level A = Administrator is not required to have advanced training in assessment and interpretation. Level B = Administrator is required to have some training in assessment and interpretation. Level C = Administrator is required to have extensive training in assessment and interpretation (e.g., a doctorate in psychology, professional licensure)

In addition to the references listed above, the following sources were used: AERA, APA, & NCME (1999); Bracken et al. (1998); CanChild Centre for Childhood Disability Research (2004); Caselman & Self (2008); Sandal, Hemmeter, Smith, & McLean (2005); Scientific Advisory Committee on Medical Outcomes Trust (2002); and Tyson & Connell (2009).

Table 3. *Measures Evaluated against Criteria for Technical Adequacy and Usability*

	Reliability			Validity					Standardization			Usability			
Measures	IC	IR	TR	Cn	Cv	Cr	Co	SS	ECS	SG	St	RL	AT	TL	AL
							Assessment								
ABAS-II	3	2	3	3	2	3	3	NE	3	3	3	3	3	1	3
BASC-2	3	2	3	3	2	NE	3	NE	3	3	3	3	3	1	1
CBCL/ 1.5-5	3	1	2	3	2	3	3	NE	3	NE	3	3	1	1	1
ITSEA	3	3	2	3	1	3	3	NE	3	3	3	3	1	1	1
SIB-R	3	2	2	3	3	3	NE	NE	3	3	3	2	2	2	3
Vineland-II	3	2	2	3	2	NE	3	NE	1	2	3	2	2	1	3
Vineland SEEC	3	1	2	3	1	3	3	NE	3	3	3	2	2	1	3
							Screening								
ASQ:SE	3	NE	3	3	NE	3	NE	2	3	3	3	3	3	3	3
BITSEA	NE	2	3	3	1	3	NE	3	3	3	3	3	3	3	3
Greenspan	3	NE	NE	3	NE	3	NE	NE	1	1	3	NE	3	3	3

Note. NE= no evidence; 1= criterion not met; 2= criterion partially met; 3= criterion met. See Table 2, column 1 for descriptions of criteria

Table 4. *Reliability Coefficients for the 10 SE Measures*

Measure	Internal Consistency*	Inter-rater*	Test-Retest*
		Assessment	
ABAS-II			
Parent Form	.91-.97	.72-.86	.84-.88
Teacher Form	.93-.98	.74-.87	.88- .91
BASC-2			
Parent Scale	.85-.93	.66-.84	.81-.86
Teacher Scale	.87-.96	.61-.81	.84-.87
CBCL/1.5-5			
Parent Form	.89-.95	.59-.67	.87-.90
Teacher Form	.89-.97	.64-.79	.77-.89
ITSEA	.85-.90	.72-.79	.76-.91
SIB-R			
Full form	.85-.97	NE	NE
EDF	.84	.91	.97
Vineland-II			
Parent Form	.90-.97	.60-.87	.82-.96
Teacher Form	.91-.98	.49-.59	.72-.90
Vineland SEEC	.80-.93	.47-.60	.71-.79
		Screening	
ASQ:SE	.82	NE	94% (McNemar test)
BITSEA	NE	.63-.74	82.-.92
GREENSPAN	.82-.89	NE	NE

Note: The coefficients are based on domain/scale and/or composite scores for the early childhood group; sub-domain (scale) reliabilities are not included. If the manual reported coefficients across multiple age groups with no combined group data, a median value was chosen. The early childhood sample used varied across measures by age (e.g., ages 2-4 for SIB-R full form compared to 0-5 for Vineland-II). Intervals for test-retest reliability ranged from 2 to 70 days. NE=no evidence; it was either not calculated or was not available for the early childhood age group.
*Exact values are available upon request.

Table 5. *Validity Evidence for the 10 SE Measures*

Measures	Convergent*	Criterion-Related	Construct	Sensitivity /Specificity
		Assessment		
ABAS-II				
Parent Form	.18-.70	√	√	NE
Teacher Form	.59-.84			
BASC-2				
Teacher Scale	.72-.85	NE	√	NE
Parent Scale	.69-.82			
CBCL/1.5-5 Parent form	.56 -.70	√	√	NE
Teacher form	.77			
ITSEA	.34-.69	√	√	√
SIB-R				
Full form	.96-.99	√	NE	NE
EDF	.77-.90			
Vineland-II				
Parent form	.46-.95	NE	√	NE
Teacher form	.52-.87			
Vineland SEEC	.63-.65	√	√	NE
		Screening		
ASQ:SE	NE	√	NE	√
BITSEA	-.34 -.60	√	NE	√
Greenspan	NE	√	NE	NE

Note: The convergent validity evidence reported are based primarily on correlations indicating the relationship between the measure's composite score and/or domain/scale scores with corresponding composite and/or domain scores from other measures and/or from previous versions. Correlations with previous versions (e.g., SIB-R) of the same measure were typically high. NE = no evidence.

*Exact values are available upon request.

Table 6. *Standardization Information for the 10 SE Measures*

Measures	Sample (birth to 5)	Sample per age group/number of age groups (birth to 5)	Sample stratification variables and year of Census approximation
		Assessment	
ABAS-II			
T/DPF	750	100-150/7	Race/ethnicity, education, geographic regions, and gender; 1999
P/CF	1350	100-150/13	
BASC-2			
TRS	1050	102-326/4	Race/ethnicity, SES and geographic regions; 2001
PRS	1200	150-382/4	
CBCL/1.5-5			
CBCL	700	NE	Race/ethnicity, SES, and geographical regions No mention of year
C-TRF	1,192		
ITSEA	600	150/4	Race/ethnicity, parent education, geographical region, and gender; 2002
SIB-R	670	109-145/5	Race, and SES, geographic region, gender, and community type and size; 1990
Vineland-II			
Survey Form	304	72-108/12	Race/ethnicity, SES, geographic region, and gender; 2001
Teacher Form	452	100-194/3	
Vineland SEEC	1200	200/6	Race/ethnicity, parent education geographical region, gender, community size, and chronological age; 1980
		Screening	
ASQ:SE	3014	298-471/8	Ethnicity, SES, parent education, geographic region, and gender; 2000
BITSEA	600	150/4	Ethnicity, parent education level, and geographic region; 2002
Greenspan	456	50- 89/8	Race/ethnicity, parent education, and geographic regions; 2000

Note: All measures used estimates of the U.S. Census data.