

Initial Development and Validation of the Social Skills Improvement System—Social and Emotional Learning Brief Scales-Teacher Form

Journal of Psychoeducational Assessment
2020, Vol. 0(0) 1–16
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0734282920953240
journals.sagepub.com/home/jpa



Christopher J. Anthony¹ , Stephen N. Elliott²,
James C. DiPerna³, and Pui-Wa Lei³

Abstract

Despite the need for assessments targeting social and emotional learning (SEL) skills within multitiered systems of support, there are few brief SEL measures available. To address this need, this article describes the development of the Social Skills Improvement System—Social and Emotional Learning Brief Scales-Teacher Form (SSIS SELb-T), a significantly shorter version of the Social Skills Improvement System—Social and Emotional Learning Edition Rating Form-Teacher (SSIS SEL RF-T). Using the SSIS SEL standardization sample and item response theory, we identified efficient sets of items for each SSIS SEL RF-T scale and examined psychometric evidence for resulting scores. In general, SSIS SELb-T scales functioned well and very similarly to scores from corresponding SSIS SEL RF-T scales. One exception was the Self-Awareness scale which, similar to its long-form counterpart, exhibited several content and psychometric limitations. Results provide initial psychometric evidence for a time-efficient teacher-informant measure of SEL competencies with promise for use within multitiered systems of support.

Keywords

social and emotional learning, assessment, multitiered systems of support, item response theory

Schools are increasingly incorporating educational programming focused on teaching students how to regulate and manage their behavior, to understand and influence their emotions, and to make healthy and responsible choices. These competencies have come to be known as *social and emotional learning* (SEL), and a growing evidence base supports their inclusion in school curricula (e.g., DiPerna, Lei, Bellinger, & Cheng, 2015; Diperna, Lei, Bellinger, & Cheng, 2016; DiPerna, Lei, Cheng, Hart, & Bellinger, 2018; Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011) and multitiered systems of support (MTSS).

¹University of Florida, FL, USA

²Arizona State University, AZ, USA

³The Pennsylvania State University, PA, USA

Corresponding Author:

Christopher J. Anthony, School of Special Education, School Psychology, and Early Childhood Studies, University of Florida, 2-189 Norman Hall, Gainesville 32611, FL, USA.

Email: canthony@coe.ufl.edu

Although there are various models of SEL skills, the framework developed by the Collaborative for Academic, Social, and Emotional Learning (CASEL) has become synonymous with SEL for many practitioners and researchers. The CASEL model evolved over time as SEL researchers and advocates attempted to clarify and organize the diverse subject matter that falls under the SEL umbrella (Shriver & Weissberg, 2020). These efforts culminated in the identification of the five interrelated domains that comprise the CASEL framework. These domains include self-awareness, self-management, social awareness, relationship skills, and responsible decision-making (Table 1). Beyond offering a template for SEL-focused research and practice, this model is gaining an increasing amount of empirical support, while becoming highly influential in SEL policy. Specifically, the CASEL model has influenced SEL policies in all 50 US states and internationally (Dusenbury, Yoder, Dermody, & Weissberg, 2019; Eklund, Kilpatrick, Kilgus, & Haider, 2018). Recent studies have also corroborated the CASEL framework as a broad model of important SEL skills (Anthony, Elliott, DiPerna, & Lei, 2020a, 2020b; Doromal, Cottone, & Kim, 2019), further supporting the utility of this framework. Despite these developments, there remain important barriers to effective practice in contemporary school-based service delivery paradigms, such as MTSS.

MTSS and the Role of Assessments

Many of the evolving SEL programs in schools are situated in MTSS. Within MTSS service delivery models, it is important to consider the multiple levels of assessment and intervention (Sugai & Horner, 2009). Given differences in the number of students served and intensity of difficulties addressed within each tier, assessments should be optimized accordingly. For example, at the tertiary level, assessments should be comprehensive in construct coverage given they are intended to inform high-stakes decisions and development of intensive interventions at the individual level. Instruments such as comprehensive rating scales or direct measures of emotion recognition are appropriate at this service delivery level (McKown, 2017). At the universal level, however, brief yet psychometrically sound measures are needed to support universal screening, periodic progress monitoring, and intervention need decisions. Because these measures need to be

Table 1. Definitions of Each Collaborative for Academic, Social, and Emotional Learning Domain and Example Items.

| Domain definition | Example item |
|--|---|
| Self-awareness: The ability to accurately recognize one's emotions and thoughts and their influence on behavior | <i>Says bad things about self</i> |
| Self-management: The ability to regulate one's emotions, thoughts, and behaviors effectively in different situations | <i>Completes tasks without bothering others</i> |
| Social awareness: The ability to take the perspective of and empathize with others from diverse backgrounds and cultures to understand social and ethical norms for behavior and to recognize family, school, and community resources, and supports | <i>Shows concern for others</i> |
| Relationship skills: The ability to establish and maintain healthy and rewarding relationships with diverse individuals and groups | <i>Interacts well with other children</i> |
| Responsible decision-making: The ability to make constructive and respectful choices about personal behavior and social interactions based on consideration of ethical standards, safety concerns, social norms, the realistic evaluation of consequences of various actions, and the well-being of self and others | <i>Acts responsibly when with others</i> |

Note. All definitions directly quoted from CASEL (2015, p. 5–6).

completed multiple times or for *all* students, they must be efficient and yield information about students' strengths and weaknesses to be useful.

Despite this increased interest in SEL and the prominence of the CASEL framework, there are very few assessments that are both aligned with the CASEL framework and brief enough for use at universal tiers of multitiered support systems. This is specifically the case for teacher-reported behavior rating scales, which serve as the cornerstone for many socioemotional and behavioral screening paradigms. The CASEL Assessment Guide identifies eight teacher/staff report measures (CASEL, 2019). Of these assessments, only two are relatively brief. First, the Social Skills Improvement System SEL Edition Screening and Progress Monitoring Scales (SSIS SEL S/PM; Elliott & Gresham, 2017a) is a brief, criterion-referenced measure in which teachers complete global judgments on student competencies in each CASEL domain. Despite its strengths in terms of efficiency, the SSIS SEL-S/PM is not a traditional rating scale and only provides teacher judgments about global competency rather than information about specific behavioral indicators of SEL.

The other brief, CASEL-aligned measure is the Devereux Student Strengths Assessment-Mini (DESSA-mini; Naglieri, LeBuffe, & Shapiro, 2011/2014). This measure can be completed in 5 minutes or less, aligned with the CASEL framework, and yields norm-referenced scores for a well-defined sample of students across multiple grades. The DESSA-mini is an 8-item version of the 72-item DESSA (LeBuffe, Shapiro, & Naglieri, 2009/2014) for students in Grades K-8. The DESSA-mini is norm-referenced behavior rating scale that yields an overall Social-Emotional Composite score based on items selected from the full-length DESSA, which includes scales targeting all five CASEL domains as well as several other competency areas. Despite its strengths, the DESSA-mini has four forms with only eight items each. These items were chosen based on their correlation with the DESSA total score, and each form does not necessarily include content from each CASEL domain. Furthermore, the DESSA-mini yields a single Social-Emotional Total score, and no DESSA-mini scores are available for individual CASEL domains.

Developing Efficient, Informative SEL Assessments and the Current Study

One method to address the need for efficient SEL assessments is through the development of brief forms based on existing measures. As previously mentioned, Naglieri et al. (2011) developed the DESSA-mini by selecting eight items from the full-length DESSA such that the resulting score (the DESSA-mini) was most strongly correlated with the DESSA total score. This process favors information about students' broad SEL skills above information about specific SEL domains. Another method for developing efficient measures involves the use of item response theory (IRT; Anthony & DiPerna, 2017, 2018; Anthony, DiPerna, & Lei, 2016; Moulton, von der Embse, Kilgus, & Drymond, 2019).

Given the strengths of IRT for such applications, several recent studies have utilized IRT to identify and pilot potential short forms of various measures. For example, Anthony et al. (2016) utilized IRT to identify sets of maximally efficient items from the Social Skills Improvement System Rating Scale (SSIS-RS; Gresham & Elliott, 2008), and a follow-up study with an independent sample of 302 students provided support for the validity of scores from these sets of items (Anthony & DiPerna, 2019). Notably, the resulting form included only 58% of the total number of SSIS-RS items and could be completed in slightly over half the time required for the full-length SSIS-RS. Such reductions in time allow for more efficient use of the measures in applications that would otherwise be impractical.

In line with these advances, our goal was to develop a brief CASEL-aligned measure well suited for multiple applications within schools implementing MTSS and gather initial reliability

and validity evidence for this measure (the Social Skills Improvement System—Social and Emotional Learning Brief Scales-Teacher Form [SSIS SEL b -T]). Specifically, we were guided by the following development goals: (a) significantly reduce the length (i.e., by at least 50%) of the Social Skills Improvement System—Social and Emotional Learning Edition Rating Form-Teacher (SSIS SEL RF-T) in the creation of the SSIS SEL b -T, (b) retain appropriate content coverage of the SSIS SEL RF-T for each SSIS SEL b -T scale, (c) produce SSIS SEL b -T scales yielding scores demonstrating sufficient reliability for low-stakes decision-making (e.g., universal screening and periodic progress monitoring), and (d) produce SSIS SEL b -T scales yielding scores with strong evidence of validity.

Method

Participants

Participants were drawn from the standardization sample of the SSIS SEL RF-T (Gresham & Elliott, 2017). Although the SSIS SEL RF-T included data for preschool children, the current investigation exclusively focused on children in Grades K-12 due to the developmental differences between preschool- and school-aged children (e.g., Flook, Zahn-Waxler, & Davidson, 2019), as well as the different applied contexts of assessments in preschools versus K-12 schools (e.g., Shepley & Grisham-Brown, 2019). Demographic characteristics of the student sample ($N = 750$) featured in the current study are reported in Table 2, which also shows that the demographics of the sample were similar to the current US student population.

Measures

Social Skills Improvement System SEL Edition Rating Forms-Teacher. The SSIS SEL RF-T Rating Scale (Gresham & Elliott, 2017) is a nationally normed behavior rating scales of SEL for students aged 3–18 years. The SSIS SEL RF-T includes 51 items rated on a 4-point Likert scale from 0 (*Never*) to 3 (*Almost Always*). There is substantial evidence for reliability of SSIS SEL RF-T scores. For example, Cronbach's α ranged from .72 to .95 across the five SSIS SEL RF-T scales and the SEL composite. Furthermore, for the SSIS SEL RF-T, 144 students were rated twice by the same teacher. Stability coefficients were generally in the low .80s. Another subsample of students ($N = 54$) was rated by pairs of teachers, and interrater reliability coefficients ranged from the upper .30s to the .60s with a median reliability coefficient of .53. There also is evidence for the validity of SSIS SEL RF-T scores. For example, the SSIS SEL RF-T Composite score was moderately negatively correlated with the Behavior Assessment System for Children-Second Edition (BASC-2) (Reynolds & Kamphaus, 2004) Externalizing Problems Composite ($r = -.55$), Internalizing Problems Composite ($r = -.46$), and School Problems Composite ($r = -.68$), but strongly positively correlated with the Adaptive Skills Composite ($r = .82$). Finally, confirmatory factor analyses also provided mixed support of the internal structure of the SSIS SEL RF-T yielding a six-factor model (with an academic competence factor included in the full-length SSIS SEL RF-T).

Social Skill Rating System-Teacher Rating Scale. The Social Skills Rating System-Teacher Rating Scale (SSRS-T) (Gresham & Elliott, 1990) is a multirater assessment of students' social skills and problem behaviors. Each SSRS item is rated on a 3-point frequency scale (0 = *Never*, 1 = *Sometimes*, or 2 = *Very Often*) based on the rater's perception of the frequency of the behavior. The SSRS-T includes three social skill domains: cooperation, assertion, and self-control. The SSRS also has three problem behavior domains: externalizing, internalizing, and hyperactivity. The SSRS-T was normed on a national sample of over 3000 students from kindergarten through high school, with equal numbers of males and females in the normative sample. The SSRS

Table 2. Demographic Characteristics (Percentages) of Participants Relative to US Student Population.

| Characteristic | Sample (N = 750) | Population ^a |
|--------------------------|------------------|-------------------------|
| Female | 50 | 49 ^b |
| Race | | |
| White | 59 | 48 |
| Black | 16 | 15 |
| Hispanic | 19 | 26 |
| Other | 6 | 10 |
| Grade | | |
| K–3 | 42 | 30 |
| 4–6 | 26 | 23 |
| 7–8 | 16 | 15 |
| 8–12 | 16 | 30 |
| Region | | |
| Northeast | 18 | 16 |
| Midwest | 22 | 21 |
| South | 36 | 39 |
| West | 25 | 24 |
| Parent's education level | | |
| Grade 11 or less | 14 | 11 |
| Grade 12 or GED | 29 | 19 |
| 1–3 years of college | 31 | 26 |
| 4 + years of college | 27 | 44 |
| Educational status | | |
| General education | 92 | 87 |
| Special education | 8 | 13 |

Note. Some percentages do not sum to 100 due to rounding.

^aExcept where noted, estimates from the 2016–2017 Digest of Educational Statistics (Snyder, de Brey, & Dillow, 2019).

^bDerived from the most recent estimates from the 2013–2014 Civil Rights Data Collection survey (U.S. Department of Education Office for Civil Rights, 2016).

demonstrates excellent psychometric properties in terms of internal consistency (social skills scales mean $\alpha = .94$; problem behaviors mean $\alpha = .87$) and test–retest reliabilities (social skills scales mean $r = .85$; problem behaviors mean $r = .84$), relationships with other measures, and factor structures (see Gresham & Elliott, 1990). The SSRS manual contains a comprehensive presentation of this information.

Behavior Assessment System for Children-Second Edition. The BASC-2 (Reynolds & Kamphaus, 2004) also was used as a validity measure. The BASC-2 is a norm-referenced diagnostic tool for assessing the behavior of children and youth aged 2–25 years. The BASC-2 Teacher Rating Scales consist of 16 primary scales and five composite scales (Adaptive Skills, Behavioral Symptoms Index, Externalizing Problems, Internalizing Problems, and School Problems). The BASC-2 manual provides evidence to support the reliability of Teacher Rating Scale scores. Specifically, internal consistency coefficients are in the .90s for the composite scales for both a general sample and a clinical sample. Retest reliability with the BASC-2 Teacher 1–8 weeks after the first administration yielded average correlations in .80s for composite scores. Finally, interrater reliability analysis yielded median reliabilities for composite scores ranging from .57 to .74.

Vineland adaptive behavior scales, second edition. The Vineland-II (Sparrow, Cicchetti, & Balla, 2005) also was used as a validity measure. The Vineland-II measures adaptive behavior of

individuals from birth to age 90 years. Evidence to support the reliability of Vineland II scores is substantial. Specifically, based on a split-half reliability test using the standardization sample data across age-groups, the five domains correlations ranged from .77 to .93. Test–retest reliability over a 1-month period ($N = 414$) ranged between .74 and .98 across domains, subdomains, and ages. Finally, interrater reliability of scores from different respondents about the same individual yielded average correlations ranging from .71 to .81 across domains/subdomains. Validity evidence reported in the Vineland-II Manual (Sparrow et al., 2005) also provides support for scores. For example, with regard to the test structure, the results of confirmatory factor analyses with the standardization sample indicated that the data fit the proposed five-factor model well.

Procedure

Data for the SSIS SEL RF-T, as well as the current study, were collected as part of the original SSIS-RS standardization in 2006–2007. Pearson Assessment field staff recruited school site coordinators in 115 schools across 36 states, who in turn, recruited participants to fit demographic targets based on the 2006 Current Population Survey (U.S. Census Bureau, 2006). These site coordinators and their schools distributed and collected the rating scales from September 2006 to October 2007. Given the data collection procedures and close monitoring of materials, these steps resulted in no missing data. The final original sample was selected from the larger respondent sample to fit 2006 US Census¹ demographics of age, gender, race/ethnicity, and educational status. Subsamples of the total sample were rated by the same teacher twice for test–retest reliability calculation ($n = 144$) and by two different teachers for interrater reliability calculations ($n = 54$). Finally, teachers completed validity measures for subsamples of the total original sample for the SSRS ($n = 221$), BASC-2 ($n = 57$), and Vineland-II ($n = 47$).

Data Analysis

We conducted a series of analyses to identify, select, and validate sets of items for the SSIS SEL b -T. To accomplish these goals, data analysis proceeded in several steps.

IRT analysis and item selection. Our first set of analyses focused on selecting items for the SSIS SEL b -T. First, we checked IRT model assumptions for each SSIS SEL RF-T scale, including the assumptions of unidimensionality and local independence. We checked the assumption of unidimensionality by conducting exploratory factor analyses by SSIS SEL RF-T scale using Mplus Version 8.1 (Muthen & Muthen, 2017). In these analyses, we treated items as categorical and used the recommended robust estimator (the Weighted Least Squares Mean and Variance Adjusted estimator utilized in Mplus). We considered evidence supportive of essential unidimensionality (i.e., sufficiently unidimensional to support IRT analyses) if the ratio of the first to the second eigenvalues exceeded 4 (Reeve, Hays, Chang, & Perfetto, 2007). In cases where this criterion was not met, we eliminated the lowest loading items until essential unidimensionality was achieved. Next, we checked the assumption of local independence utilizing local dependence (LD) standardized χ^2 values output by IRTPro version 4 (Cai, Thissen, & du Toit, 2019). As recommended, values exceeding 10 were considered to indicate excessive LD (Cai et al., 2019). Finally, we checked overall model fit for the graded response model (Samejima, 1969) analysis model via the Root Mean Square Error of Approximation (RMSEA) values provided by IRTPro. Specifically, RMSEA values less than .10 (MacCallum, Browne, & Sugawara, 1996) were considered to indicate adequate fit to the model.

Next, we evaluated individual items to identify which would be retained for the SSIS SEL b -T. We considered several indications of each item's technical quality during this process. First, we used the item information functions produced by IRTPro to evaluate the precision and utility of all items. In consideration of item information functions, we focused on information provided in the

“at-risk” range (-1.5 to $-.5$ on the θ scale; Anthony et al., 2016) given the likely intended use of the SSIS SEL*b*-T. Our goal for each scale was to ensure that the Test Information Function (TIF) exceeded 5 across the “at-risk” range on the θ scale. This criterion corresponds to a .80 reliability criterion (a commonly used criterion for screening and low-stakes decisions; Salvia, Ysseldyke, & Witmer, 2016) using a formula to convert information to a more familiar reliability metric ($1 - [1/\text{information}]$; Petrillo, Cano, McLeod, & Coon, 2015).

We also considered LD during item selection to avoid selecting items that would result in the inclusion of any item pair with standardized LD χ^2 values greater than 10 (Cai et al., 2019). Finally, we considered whether items displayed evidence of differential item functioning (DIF) by utilizing the two-step DIF purification procedure outlined by Tay, Meade, and Cao (2015) to conduct DIF analysis for gender (male vs. female) and race (white vs. nonwhite). To further probe DIF, we calculated effect size for items with statistically significant DIF violations. Specifically, we calculated the expected score standardized difference (ESSD) between focal and reference groups based on IRT parameter estimates using Visual DF (Meade, 2010). The ESSD is expressed in *SD* units and thus can be interpreted according to standard criteria for Cohen’s *d* (i.e., .2 = small; .5 = medium; .8 = large) as suggested by Meade.

After we compiled these indications of items’ technical quality, a document was generated by the first author summarizing all information for each SSIS SEL*b*-T scale. This document was distributed to each author, who independently reviewed this information and considered which items to include on the final forms, ultimately identifying four to six items for inclusion (our original goal was to include four items per scale for a total scale length of 20 items). After independent review, the authors discussed item selection considerations, including the psychometric information for each item (e.g., LD, DIF, and item information) as well as content considerations (e.g., alignment with the CASEL framework) until consensus was reached. This process was conducted iteratively until each SSIS SEL*b*-T scale was identified.

Examination of initial reliability and validity evidence. After identifying our final set of items for the SSIS SEL*b*-T, we conducted initial reliability and validity analyses on scores from our identified SSIS SEL*b*-T scales. Our primary indicator of score reliability was the TIFs produced by IRTPro. We plotted curves for each SSIS SEL*b*-T scale and then considered the level of information produced by each scale with a special focus on the at-risk range (-1.5 to $-.5$ on the latent trait scale). In addition to TIFs, we also computed Cronbach’s α , test–retest reliability coefficients, and interrater reliability coefficients for each SSIS SEL*b*-T scale. Each of these indices also were computed for the full-length SSIS SEL RF-T for comparison purposes. To examine validity of the SSIS SEL*b*-T scale scores, we computed validity coefficients for each SSIS SEL*b*-T scale with the SSRS-T, BASC-2, and Vineland-II. We computed corresponding coefficients for each SSIS SEL RF-T scale and compared them to SSIS SEL*b*-T correlations using Steiger’s (1980) formula.

Results

IRT Assumptions and DIF

First, we conducted analyses informing the selection of SSIS SEL*b*-T items, including checking IRT assumptions and evaluating DIF. Initially, ratios of first to second eigenvalues ranged from 2.86 to 8.43 (median = 4.92), and two scales did not meet our a priori criterion (self-awareness and responsible decision-making). After excluding the lowest loading item on each of these scales, our a priori criterion was met in all cases, with ratios of first to second eigenvalues ranging from 4.44 to 8.43 (median = 4.92). With regard to model fit, RMSEA values ranged from .03 to .07 (median = .06) across SSIS SEL RF-T scales. We also computed LD χ^2 values. Percentages of item pairs evidencing LD ranged from 0 to 29% (median = 13%) across scales.

DIF analyses were conducted to inform final item selection. On most scales, there were only a few instances of items that showed evidence of DIF. Specifically, there were six instances of gender-based DIF and six instances of race-based DIF. Most of these instances involved the same four items on the Self-Awareness scale, which evidenced both gender- and race-based DIF. We considered this evidence, LD values, item and TIFs, and content coverage during item selection. This process led to the identification of 20 items for the SSIS SEL*b*-T. These items had no LD violations, two gender-based DIF violations (one on the Self-Awareness scale and one on the Social Awareness scale), and one race-based DIF violation (on the Self-Awareness scale). The DIF item was retained for the Self-Awareness scale because it had lower ESSD values (ESSD = $-.56$ for race DIF and 1.01 for gender DIF) than any other available item. The single item retained with gender-based DIF for the Social Awareness scale was associated with a small DIF effect (ESSD = $-.37$) and is expected to have minimum effects on the overall scale. After selecting items, we evaluated each SSIS SEL*b*-T scale according to our development goals for the measure.

Scale Length and Content Coverage

First, we evaluated the extent to which the SSIS SEL*b*-T resulted in a more efficient measure than the SSIS SEL RF-T and the content validity of the scale. In line with our first development goal, the application of the process of item selection resulted in forms that were much shorter than the original SSIS SEL RF-T. Specifically, 31 items of the original 51 SSIS SEL RF-T items were pruned in the development of the 20-item SSIS SEL*b*-T rendering the measure 61% shorter. Based on the reported SSIS SEL RF-T completion time of 10 minutes (Gresham & Elliott, 2017), the SSIS SEL*b*-T should be able to be completed in under 5 minutes.

Next, in line with our second development goal, our process of conducting analyses on a scale-by-scale basis ensured that each of the five CASEL domains was represented in the SSIS SEL*b*-T. Despite this, the content validity of the original SSIS SEL RF-T item pools did vary by domain, resulting in some unevenness in content validity across SSIS SEL*b*-T scale. Specifically, the SSIS SEL RF-T Self-Awareness scale consists primarily of reverse-coded items reflecting Internalizing Problems, rendering the corresponding SSIS SEL*b*-T scale as likely the least CASEL aligned of scales on the SSIS SEL*b*-T.

Reliability Evidence

In line with our third development goal, we evaluated reliability evidence for each SSIS SEL*b*-T scale as well as the SSIS SEL*b*-T Composite (Table 3). Cronbach's α was .93 for the SSIS SEL*b*-T Composite and ranged from .79 to .87 (median = .83) across SSIS SEL*b*-T scales. The test-retest reliability coefficient was .84 for the SSIS SEL*b*-T Composite, and these coefficients ranged from .75 to .83 across SSIS SEL*b*-T scales (median = .78). Finally, the interrater reliability coefficient was .65 for the SSIS SEL*b*-T Composite and ranged from .47 to .65 (median = .51) across SSIS SEL*b*-T scales. With regard to SSIS SEL*b*-T scale TIFs, in general, scale information exceeded the five (.80 reliability) a priori criterion across a wide spectrum of the targeted constructs (see Figure 1).

Validity Evidence

Finally, to inform our evaluation of our fourth development goal, evidence for the validity of SSIS SEL*b*-T scores was determined from several sources. First, score intercorrelations were largely as expected (Table 4) and ranged from .28 to .86 (median = .66). One scale in particular, Self-Awareness had intercorrelations of smaller magnitude (.28-.38; median = .31) than other SSIS

Table 3. SSIS SEL Brief Scales-Teacher and SSIS SEL Edition Rating Form-Teacher Reliability Statistics.

| Scale | Cronbach's α | | Test-retest | | Interrater | |
|-----------------------------|---------------------|-----|-------------|-----|------------|-----|
| | (n = 750) | | (n = 144) | | (n = 54) | |
| | SELb | SEL | SELb | SEL | SELb | SEL |
| Self-awareness | .83 | .80 | .83 | .81 | .47 | .51 |
| Self-management | .83 | .91 | .83 | .79 | .65 | .67 |
| Responsible decision-making | .87 | .81 | .78 | .83 | .51 | .51 |
| Relationship skills | .79 | .90 | .75 | .78 | .58 | .71 |
| Social awareness | .83 | .91 | .75 | .86 | .47 | .56 |
| SEL composite | .93 | .96 | .84 | .85 | .65 | .71 |

Note. SSIS = Social Skills Improvement System; SELb = Social Skills Improvement System Social and Emotional Learning Edition Brief Scales-Teacher; SEL = Social Skills Improvement System Social and Emotional Learning Edition Rating Form-Teacher. All correlations statistically significant ($p < .001$).

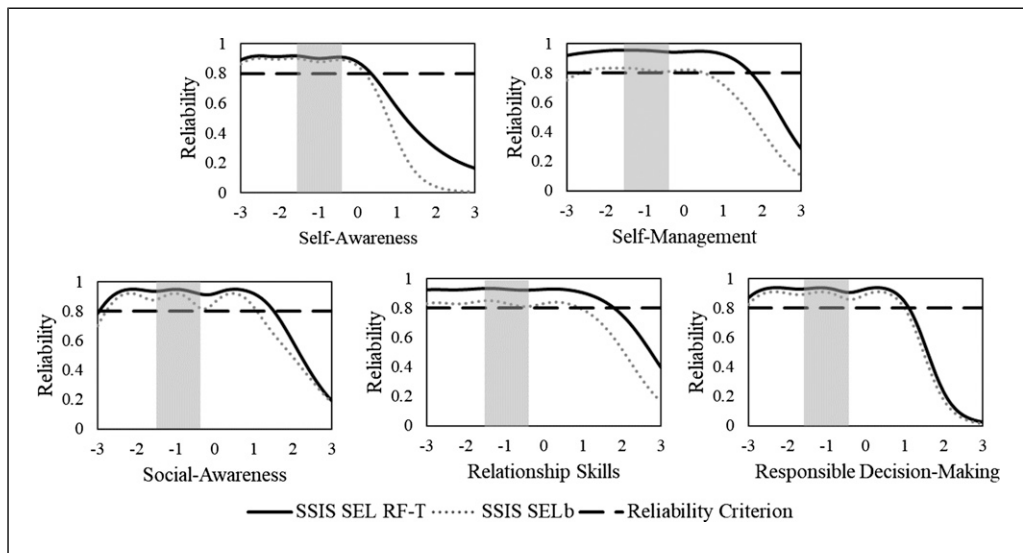


Figure 1. Test information functions for Social Skills Improvement System Social and Emotional Learning Edition Rating Form-Teacher (SSIS SEL RF-T) and SSIS SEL Brief Edition-Teacher Form (SSIS SELb-T). Note. Reliability on y-axis converted from total information with the following formula: $1 - (1/\text{information})$ as recommended by [Petrillo et al. \(2015\)](#). Shaded region represents the “at-risk” range.

SELb-T scales. With regard to convergent validity evidence, the pattern of SSIS SELb-T score correlations with SSRS-T, BASC-2, and Vineland-II scores ([Tables 5 and 6](#)) generally was as expected. Specifically, correlations with the SSRS Social Skills scales ranged from .49 to .71 (median = .59) for the SSIS SELb-T Composite and from .22 to .68 (median = .49) for the SSIS SELb-T scales. Corresponding correlations for the SSRS Problem Behaviors scales ranged from $-.63$ to $-.40$ (median = $-.56$) for the SSIS SELb-T Composite and from $-.73$ to $-.20$ (median = $-.38$) for the SSIS SELb-T scales. Next, SSIS SELb-T correlations with the BASC-2 Externalizing, Internalizing, School Problems, and Behavioral Symptoms Index scores ranged from $-.81$ to $-.58$ (median = $-.72$) for the SSIS SELb-T Composite and from $-.82$ to $-.29$ (median = $-.64$) for the SSIS SELb-T scales. The correlation with the BASC-2 Adaptive Behavior Index score was .85 for the SSIS SELb-T

Table 4. SSIS SEL Brief Scales-Teacher and SSIS SEL Edition Rating Form-Teacher Intercorrelations and 95% Confidence Intervals.

| | 1 | 2 | 3 | 4 | 5 | 6 |
|--------------------------------------|----------------|----------------|----------------|----------------|----------------|---|
| SSIS SEL brief scales-teacher | | | | | | |
| 1. Self-awareness | — | | | | | |
| 2. Self-management | .28 [.21, .34] | — | | | | |
| 3. Responsible decision-making | .33 [.27, .39] | .86 [.84, .88] | — | | | |
| 4. Relationship skills | .38 [.32, .44] | .71 [.67, .74] | .77 [.74, .80] | — | | |
| 5. Social awareness | .28 [.21, .34] | .63 [.59, .67] | .69 [.65, .73] | .75 [.72, .78] | — | |
| 6. SEL composite | .53 [.48, .58] | .87 [.85, .89] | .91 [.90, .92] | .89 [.88, .90] | .83 [.81, .85] | — |
| SSIS SEL edition rating form-teacher | | | | | | |
| 1. Self-awareness | — | | | | | |
| 2. Self-management | .51 [.46, .56] | — | | | | |
| 3. Responsible decision-making | .48 [.42, .53] | .85 [.83, .87] | — | | | |
| 4. Relationship skills | .59 [.54, .63] | .77 [.74, .80] | .79 [.76, .82] | — | | |
| 5. Social awareness | .38 [.32, .44] | .74 [.71, .77] | .73 [.70, .76] | .76 [.73, .79] | — | |
| 6. SEL composite | .66 [.62, .70] | .93 [.92, .94] | .90 [.87, .91] | .92 [.91, .93] | .84 [.82, .86] | — |

Note. All correlations statistically significant ($p < .001$). SSIS SEL = Social Skills Improvement System Social and Emotional Learning.

Composite and ranged from .62 to .82 (median = .75) for the SSIS SEL_b-T scales. Finally, with regard to Vineland-II scores, correlations between the SSIS SEL_b-T Composite and Vineland-II Composite scores ranged from .62 to .72 (median = .69) and corresponding correlations between SSIS SEL_b-T scale scores and Vineland-II Composites ranged from .31 to .72 (median = .58).

Discussion

The purpose of this study was to develop and examine initial validity evidence for the SSIS SEL_b-T, a brief, time-efficient, CASEL-aligned rating scale to meet assessment needs within MTSS and Positive Behavioral Interventions and Supports. The results of these efforts largely were successful, although there are important areas for further development. Specifically, psychometric evidence was strong for four of the SSIS SEL_b-T scales. Estimates showed that SSIS SEL_b-T scores generally met reliability standards for the type of low-stakes decisions the SSIS SEL_b-T would be best used (e.g., initial universal screening; [Salvia et al., 2016](#)). This level of precision supports fairly broad applicability of the SSIS SEL_b-T for screening and progress monitoring applications within MTSS, as well as for use in research where efficient, sensitive assessments are valued.

With regard to validity evidence, SSIS SEL_b-T scores were generally correlated with each other and with scores from other measures as expected. In addition, scores from the SSIS SEL_b-T functioned very similarly to corresponding scores from the full-length SSIS SEL Rating Form-Teacher with regard to both reliability and validity. The only domain in which validity evidence appeared to differ slightly was for scale intercorrelations between Self-Awareness scores and other SEL scales. Thus, overall, these sources of evidence augment the overall reliability and validity evidence and indicate the SSIS SEL_b-T likely functions very similarly to the SSIS SEL RF-T.

Limitations and Future Research Directions

Despite generally promising evidence, there were important limitations both with the study as well as with the SSIS SEL RF-T and SSIS SEL_b-T teacher forms. First, although the sample was larger and more demographically representative relative to similar IRT projects that have used

Table 5. SSIS SEL Brief Scales-Teacher and SSIS SEL Edition Rating Form-Teacher Validity Coefficients with Social Skills Rating System (N = 221).

| Scale | Self-awareness | | Self-management | | Social awareness | | Relationship skills | | Responsible decision-making | | SEL composite | |
|----------------------------|----------------|------------|-----------------|-------------|------------------|-------------|---------------------|------------|-----------------------------|-------------|---------------|-------------|
| | SELb | SEL | SELb | SEL | SELb | SEL | SELb | SEL | SELb | SEL | SELb | SEL |
| Social skills | | | | | | | | | | | | |
| Cooperation | .23 | .26 | .63 | .59 | .39 | .37 | .47 | .46 | .63 | .57 | .59 | .55 |
| Assertion | .40 | .52 | .22 | .37 | .49 | .49 | .52 | .61 | .33 | .47 | .49 | .57 |
| Self-control | .39 | .45 | .60 | .68 | .56 | .55 | .68 | .62 | .62 | .60 | .71 | .69 |
| Problem behaviors | | | | | | | | | | | | |
| Externalizing | -.36 | -.37 | -.55 | -.58 | -.38 | -.36 | -.43 | -.38 | -.52 | -.42 | -.56 | -.51 |
| Internalizing | -.69 | -.69 | -.20 | -.31 | -.20 | -.18 | -.35 | -.40 | -.21 | -.24 | -.40 | -.41 |
| Hyperactivity ^a | -.35 | -.41 | -.73 | -.69 | -.36 | -.32 | -.42 | -.42 | -.65 | -.58 | -.63 | -.58 |

Note. SSIS = Social Skills Improvement System; SELb = Social Skills Improvement System Social and Emotional Learning Edition Brief Scales-Teacher; SEL = Social Skills Improvement System Social and Emotional Learning Edition Rating Form-Teacher; Bolded coefficients indicate statistically significant differences between validity coefficients.

^aOnly completed for elementary-aged children (n = 146).

Table 6. SSIS SEL Brief Scales-Teacher and SSIS SEL Edition Rating Form-Teacher Validity Coefficients.

| | Self-awareness | | Self-management | | Social awareness | | Relationship skills | | Responsible decision-making | | SEL composite | |
|---|----------------|------------|-----------------|------------|------------------|------------|---------------------|------------|-----------------------------|------------|---------------|------------|
| | SELb | SEL | SELb | SEL | SELb | SEL | SELb | SEL | SELb | SEL | SELb | SEL |
| Behavior assessment system for children-2 | | | | | | | | | | | | |
| Externalizing | -.55 | -.53 | -.67 | -.61 | -.36 | -.35 | -.57 | -.45 | -.70 | -.56 | -.67 | -.56 |
| Internalizing | -.67 | -.71 | -.50 | -.52 | -.29 | -.26 | -.53 | -.51 | -.51 | -.49 | -.58 | -.55 |
| School problems | -.64 | -.71 | -.78 | -.75 | -.39 | -.39 | -.64 | -.63 | -.78 | -.70 | -.76 | -.72 |
| Adaptive skills | .67 | .75 | .75 | .84 | .63 | .63 | .78 | .83 | .82 | .80 | .85 | .87 |
| Behavioral symptoms | -.82 | -.79 | -.71 | -.73 | -.45 | -.46 | -.71 | -.66 | -.76 | -.67 | -.81 | -.74 |
| Vineland-II | | | | | | | | | | | | |
| Communication | .40 | .48 | .60 | .67 | .46 | .50 | .59 | .58 | .64 | .65 | .67 | .68 |
| Daily living skills | .31 | .42 | .64 | .72 | .57 | .58 | .62 | .61 | .72 | .77 | .72 | .73 |
| Socialization | .41 | .51 | .45 | .57 | .48 | .53 | .63 | .60 | .54 | .62 | .62 | .65 |
| Adaptive behavior | .40 | .50 | .58 | .68 | .52 | .55 | .65 | .62 | .66 | .72 | .70 | .72 |

Note. SSIS = Social Skills Improvement System; SELb = Social Skills Improvement System Social and Emotional Learning Edition Brief Scales-Teacher; SEL = Social Skills Improvement System Social and Emotional Learning Edition Rating Form-Teacher; Subsamples of the total standardization sample used for BASC-2 ($n = 57$) and Vineland-II ($n = 47$) correlations. Bolded coefficients indicate statistically significant differences between validity coefficients.

convenience samples (e.g., Anthony et al., 2016), data for this project were collected in 2006. Nevertheless, there is evidence that the behaviors assessed by the SSIS SEL*b*-T continue to be highly relevant in contemporary schools (e.g., Wollersheim, Sandilos, DiPerna, & Lei, 2017). Findings of the study also underscored an important limitation of the content of the self-awareness items from the SSIS SEL*b*-T and the original SSIS SEL RF-T. Specifically, scale intercorrelations were lower for the SSIS SEL*b*-T Self-Awareness scale than for any other scale by a wide margin, and DIF violations were most severe for this scale. On the original SSIS SEL RF-T, the Self-Awareness scale comprises two positively worded items and six reverse-coded items from the SSIS-RS Internalizing scale. The SSIS authors (Gresham & Elliott, 2017, p. 12) hypothesized a close relationship between internalizing behaviors and self-awareness and thus used some internalizing items to create their Self-Awareness scale. In our analyses, the two positively worded items had to be removed to either establish essential unidimensionality or provided very little information across the latent trait range indicating that most of the construct variance for this scale is explained by the reverse-coded internalizing items. As a result, the four items chosen for the SSIS SEL*b*-T reflect internalizing symptomatology, which renders the brief scale narrowly focused on indirect (and negative) indicators of self-awareness. The content of the Self-Awareness scale (on both the SSIS SEL*b*-T and SSIS SEL RF-T versions) is much less aligned with CASEL definitions than other SSIS SEL*b*-T scales. As such, future research should focus on improving the content of the SSIS SEL*b*-T Self-Awareness scale. Until such development occurs, researchers and practitioners should be cautious when using the Self-Awareness scale of both the SSIS SEL*b*-T and SSIS SEL RF-T due to these concerns.

Additionally, future research should also gather reliability and validity data for the SSIS SEL*b*-T administered as a standalone measure (Smith, McCarthy, & Anderson, 2000). The current evidence is promising, but evidence from studies administering the SSIS SEL*b* as a standalone measure would provide even stronger support for the psychometric quality of SSIS SEL*b* scores. Finally, future studies should aim to gather evidence for the utility of SSIS SEL*b*-T scores for particular applications. For example, predictive validity studies establishing cut-scores for prediction of relevant criterion outcomes would support the use of the SSIS SEL*b*-T for universal screening. Likewise, studies examining the change sensitivity of SSIS SEL*b*-T scores would contribute to evidence regarding its use in evaluating intervention outcomes.

Implications

Given the varied assessment needs in MTSS and options available, the development of efficient CASEL-aligned assessments is necessary to advance SEL-focused practice and research. The SSIS SEL*b*-T represents a 60% reduction in items relative to the SSIS SEL RF-T and can be completed in 5 minutes by teachers. Thus, the SSIS SEL*b*-T joins the DESSA-mini and the SSIS SEL-S&PM as the only teacher-report SEL measures able to be completed in 5 minutes or less per student. Unlike the other measures, however, the SSIS SEL*b*-T assesses specific SEL skills and behaviors rather than global teacher judgments and provides five scores aligned with the prominent CASEL framework. The SSIS SEL*b*-T also has other notable strengths, such as alignment with the SSIS SEL Classwide Intervention Program (CIP; Elliott & Gresham, 2017b). Thus, the SSIS SEL*b*-T holds promise for many applications including initial screening for strengths and weaknesses, periodic progress monitoring, and intervention selection for the SSIS SEL CIP. Given the lack of technically sound, time-efficient SEL measures, this shorter version of a prominent SEL assessment broadens the horizons of SEL-focused data-based decision-making within MTSS and for SEL researchers working with time-challenged teachers.

Conclusion

The lack of SEL-focused assessments has constrained the burgeoning interest in promoting SEL skills in schools (McKown, 2017). Despite high levels of attention, there remain few suitable measures, especially considering the variety of assessments needed within school contexts. The results of this study indicate that the SSIS SELb-Teacher form holds promise to advance SEL-focused assessment in schools and addresses researchers' needs for efficient assessment. The SSIS SELb-T was developed for applications that require brief, but informative, assessments for a relatively large number of students. The development of such brief measures that reflect student and parent perspectives represents important next steps in expanding the number of brief assessments for use within MTSS.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The SSIS SEL Brief Scales–Teacher Form is published by SAIL-CoLab, and all authors receive financial remuneration from the distribution of these measures.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Christopher J. Anthony  <https://orcid.org/0000-0003-1883-8322>

Note

1. This sample was also subsequently confirmed to represent the 2014 U.S. Census well during the SSIS SEL development.

References

- Anthony, C. J., & DiPerna, J. C. (2017). Identifying sets of maximally efficient items from the academic competence evaluation scales-teacher form. *School Psychology Quarterly, 32*, 552-559. doi:10.1037/spq0000205
- Anthony, C. J., & DiPerna, J. C. (2018). Piloting a short form of the academic competence evaluation scales. *School Mental Health, 10*(3), 314-321. doi:10.1007/s12310-018-9254-7
- Anthony, C. J., & DiPerna, J. C. (2019). Examining the psychometric properties of maximally efficient items from the social skills improvement system–teacher rating scale. *Journal of Psychoeducational Assessment, 37*, 307-319. doi:10.1177/0734282917743335
- Anthony, C. J., DiPerna, J. C., & Lei, P.-W. (2016). Maximizing measurement efficiency of behavior rating scales using item response theory: An example with the social skills improvement system–teacher rating scale. *Journal of School Psychology, 55*, 57-69. doi:10.1016/j.jsp.2015.12.005
- Anthony, C. J., Elliott, S. N., DiPerna, J. C., & Lei, P.-W. (2020a). The SSIS SEL brief scales–student form: Initial development and validation. *School Psychology, 35*(4)277-283. doi:10.1037/spq0000390
- Anthony, C. J., Elliott, S. N., DiPerna, J. C., & Lei, P.-W. (2020b). Multirater assessment of young children's social and emotional learning via the SSIS SEL brief scales–preschool forms. *Early Childhood Research Quarterly, 53*, 625-637. doi:10.1016/j.ecresq.2020.07.006
- Cai, L., Thissen, D., & du Toit, S. H. C. (2019). *IRTPRO for Windows [computer software]*. Lincolnwood, IL: Scientific Software International.
- CASEL. (2015). *Effective Social and Emotional Learning Programs: Middle and High School*. Chicago, IL: Author.
- CASEL. (2019). *CASEL Assessment Guide*. Retrieved from <https://measuringSEL.casel.org/access-assessment-guide/>
- DiPerna, J. C., Lei, P., Bellinger, J., & Cheng, W. (2015). Efficacy of the social skills improvement system classwide intervention program (SSIS-CIP) primary version. *School Psychology Quarterly, 30*, 123. doi:10.1037/e615512013-001

- Diperna, J. C., Lei, P., Bellinger, J., & Cheng, W. (2016). Effects of a universal positive classroom behavior program on student learning. *Psychology in the Schools, 53*, 189-203. doi:10.1002/pits.21891
- DiPerna, J. C., Lei, P., Cheng, W., Hart, S. C., & Bellinger, J. (2018). A cluster randomized trial of the social skills improvement system-classwide intervention program (SSIS-CIP) in first grade. *Journal of Educational Psychology, 110*, 1-16. doi:10.1037/edu0000191
- Doromal, J. B., Cottone, E. A., & Kim, H. (2019). Preliminary validation of the teacher-rated DESSA in a low-income, kindergarten sample. *Journal of Psychoeducational Assessment, 37*, 40-54. doi:10.1177/0734282917731460
- Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82*, 405-432. doi:10.1111/j.1467-8624.2010.01564.x
- Dusenbury, L., Yoder, N., Dermody, C., & Weissberg, R. (2019). An examination of frameworks for social and emotional learning (SEL) reflected in state K-12 learning standards. Establishing Practical Social-Emotional Competence Assessments Work Group: *Collaborative for Academic, Social, and Emotional Learning*. Retrieved from, <https://measuringSEL.caseli.org/wp-content/uploads/2019/02/Framework-C.3.pdf>
- Eklund, K., Kilpatrick, K. D., Kilgus, S. P., & Haider, A. (2018). A systematic review of state-level social-emotional learning standards: Implications for practice and research. *School Psychology Review, 47*, 316-326. doi:10.17105/spr-2017.0116.v47-3
- Elliott, S. N., & Gresham, F. M. (2017a). *Social skills improvement system SEL edition – screening/progress monitoring scales*. San Antonio, TX: Pearson Assessment.
- Elliott, S. N., & Gresham, F. M. (2017b). *Social skills improvement system – social and emotional learning edition: Classwide intervention program*. Minneapolis, MN: Pearson, Assessment.
- Flook, L., Zahn-Waxler, C., & Davidson, R. J. (2019). Developmental differences in prosocial behavior between preschool and late elementary school. *Frontiers in Psychology, 10*, 876. doi:10.3389/fpsyg.2019.00876
- Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system*. Circle Pines, MN: AGS.
- Gresham, F. M., & Elliott, S. N. (2008). *Social skills improvement system*. San Antonio, TX: Pearson.
- Gresham, F. M., & Elliott, S. N. (2017). *Social skills improvement system social emotional learning edition rating forms*. Bloomington, MN: Pearson Assessments.
- Gresham, F. M., Elliott, S. N., Metallo, S., Byrd, S., Wilson, E., Erickson, M., ... Altman, R. A. (2018). Psychometric fundamentals of the social skills improvement system social learning edition rating forms. *Assessment for Effective Intervention, 45*, 194-209. doi:10.1177/1534508418808598
- LeBuffé, P. A., Shapiro, V. B., & Naglieri, J. A. (2009/2014). *The Devereux student strengths assessment (DESSA): Assessment, technical manual, and user's guide*. Charlotte, NC: Apperson, Inc.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130. doi:10.1037/1082-989x.1.2.130
- McKown, C. (2017). Social-emotional assessment, performance, and standards. *The Future of Children, 27*, 157-178. doi:10.1353/foc.2017.0008
- Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*(4), 728-743. doi:10.1037/a0018966
- Moulton, S., von der Embse, N., Kilgus, S., & Drymond, M. (2019). Building a better behavior progress monitoring tool using maximally efficient items. *School Psychology, 34*, 695-705. doi:10.1037/spq0000334
- Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed). Los Angeles, CA: Muthén & Muthén.
- Naglieri, J. A., LeBuffé, P. A., & Shapiro, V. B. (2011/2014). *The Devereux student strengths assessment – mini (DESSA-Mini): Assessment, technical manual, and user's guide*. Charlotte, NC: Apperson.
- Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research, 18*, 25-34. doi:10.1016/j.jval.2014.10.005
- Reeve, B. B., Hays, R. D., Chang, C.-H., & Peretto, E. M. (2007). Applying item response theory to enhance health outcomes assessment. *Quality of Life Research, 16*, 1-3. doi:10.1007/s11136-007-9220-6

- Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior assessment system for children—second edition (BASC-2)*. Bloomington, Mn: Pearson.
- Salvia, J., Ysseldyke, J. E., & Witmer, S. (2016). *Assessment in special and inclusive education* (13th ed.). Boston, MA: Cengage Learning.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores, *Psychometrika*, 34, 1-97. doi:10.1007/bf03372160
- Shepley, C., & Grisham-Brown, J. (2019). Multi-tiered systems of support for preschool-aged children: A review and meta-analysis. *Early Childhood Research Quarterly*, 47, 296-308. doi:10.1016/j.ecresq.2019.01.004
- Shriver, T. P., & Weissberg, R. P. (2020). A response to constructive criticism of social and emotional learning. *Phi Delta Kappan*, 101(7), 52-57. doi:10.1177/0031721720917543
- Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, 12, 102-111. doi:10.1037/1040-3590.12.1.102
- Snyder, T. D., de Brey, C., and Dillow, S. A. (2019). *Digest of education statistics 2017 (NCES 2018-070)*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.
- Sparrow, S. S., Cicchetti, D., & Balla, D. A. (2005). *Vineland adaptive behavior scales* (2nd ed.). Circle Pines, MN: American Guidance Service.
- Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87, 245-251. doi:10.1037/0033-2909.87.2.245
- Sugai, G., & Horner, R. H. (2009). Responsiveness-to-intervention and school-wide positive behavior supports: Integration of multi-tiered system approaches. *Exceptionality*, 17, 223-237. doi:10.1080/09362830903235375
- Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, 18, 3-46. doi:10.1177/1094428114553062
- U.S. Census Bureau. (2006). *Current population survey, March 2006* [Machine-readable data file]. Washington, DC: Author.
- U.S. Department of Education Office for Civil Rights. (2016). *Civil rights data collection: A first look*. Retrieved from <https://www2.ed.gov/>
- Wollersheim, S. S., Sandilos, L. E., DiPerna, J. C., & Lei, P.-W. (2017). Social validity of the social skills improvement system-classwide intervention program (SSIS-CIP) in the primary grades. *School psychology quarterly: The official journal of the Division of School Psychology, American Psychological Association*, 32(3), 414-421.