Early Childhood Research Quarterly 53 (2020) 625–637

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

Multirater assessment of young children's social and emotional learning via the SSIS SEL Brief Scales – Preschool Forms

Christopher J. Anthony[a,*], Stephen N. Elliott[b], James C. DiPerna[c], Pui-Wa Lei[c]

[a] University of Florida, USA
[b] Arizona State University, USA
[c] The Pennsylvania State University, USA

ARTICLE INFO

Article history:
Received 17 January 2020
Received in revised form 24 June 2020
Accepted 17 July 2020

Keywords:
Social and emotional learning
Early childhood
Multi-informant assessment
Item response theory
Fairness

ABSTRACT

Interest in social and emotional learning (SEL) skills is growing rapidly with all 50 states adopting SEL standards for preschool children. However, data on these types of skills for young children are limited due to a paucity of psychometrically-sound assessments. Further, most available assessments are lengthy and minimally aligned with widely used SEL frameworks such as the model proposed by the Collaborative for Academic, Social, and Emotional Learning (CASEL). Thus, the current study focused on the development of valid and time-efficient rating scales of young children's SEL skills using teachers and parents as informants. We used item response theory to select items from the SSIS Social Emotional Learning Edition (SSIS SEL; Gresham & Elliott, 2017) using the national standardization sample of the measure. We then examined initial evidence of score reliability, validity, and fairness for the SSIS SEL Brief Scales – Preschool Forms resulting from this process. Results provide initial evidence for score reliability, validity, and fairness for both the Teacher and Parent versions of this measure.

1. Introduction

"The development of children's social and emotional learning skills has been a critical aspect, and in many cases the central focus, of early childhood education. Social and emotional learning (SEL) has been defined as a "process of acquiring knowledge, skills, attitudes, and beliefs to identify and manage emotions; to care about others; to make good decisions; to behave ethically and responsibly; to develop positive relationships and to avoid negative behaviors" (Elias & Moceri, 2012, p. 424). Common SEL skills such as attending to instructions, taking turns, following instructions, and understanding one's own and others' emotions are highly valued and needed for school readiness (e.g., Bierman, Greenberg, & Abenavoli, 2016; Denham, Bassett, Zinsser, & Wyatt, 2014). These and other skills are receiving renewed attention from early childhood educators as researchers have demonstrated their importance for dealing successfully with social and academic challenges (e.g., Denham et al., 2014).

1.1. SEL competency framework

Many conceptual frameworks exist for identifying important SEL skills domains (e.g., Jones, Bailey, Brush, & Nelson, 2019), but one in particular has gained traction in the early childhood community. Specifically, the Collaborative for Academic, Social, and Emotional Learning (CASEL) has advanced a theoretical framework of SEL, often referred to as the "CASEL Five" (CASEL, 2015), which includes: *Self-Awareness,* the ability to accurately recognize one's emotions and thoughts and their influence on behavior; *Self-Management*, the ability to regulate one's emotions, thoughts, and behaviors effectively in different situations; *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; *Relationship Skills*, the ability to establish and maintain healthy and rewarding relationships with diverse individuals and groups; and *Responsible Decision-Making Skills,* the ability to make constructive and respectful choices about personal behavior based on consideration of ethical standards, safety concerns, social norms, evaluation of consequences of various actions, and the well-being of self and others.

Although more than a dozen other SEL competency frameworks exist (Jones, Bailey, Brush, & Nelson, 2018), the CASEL framework is the most pervasive, directly influencing educational policy and the

* Corresponding author at: School of Special Education, School Psychology, & Early Childhood Studies, University of Florida, 2-189 Norman Hall, Gainesville, FL 32611, USA.
E-mail address: canthony@coe.ufl.edu (C.J. Anthony).

https://doi.org/10.1016/j.ecresq.2020.07.006
0885-2006/

development of dozens of school-based intervention programs in the United States, England, New Zealand, and Australia. For example, an examination of the CASEL State Scan report (Dusenbury, Dermody, & Weissberg, 2018) documented that all 50 states, along with the District of Columbia and five U.S. territories, have identified Pre-K competencies/standards for SEL. Furthermore, most of these state standards align very closely, if not completely, with the CASEL framework. Specifically, in a study evaluating the content alignment of Pre-K state standards with the CASEL framework, Eklund, Kilpatrick, Kilgus, and Haider (2018) concluded that 34 states and the District of Columbia included all five CASEL domains, 14 states identified four of the CASEL domains, and the remaining states included three. Based on their review, Eklund et al. concluded that the CASEL framework could serve a unifying function for promoting SEL-focused service provision similar to the function that the "Big Five" reading competencies of the National Reading Panel (2000) served for reading assessment and instruction.

1.2. Assessment of early childhood SEL competencies and skills

To promote attention to children's SEL skills and development, several direct measures of SEL-related constructs have been developed and refined over the past few decades. Indeed, Denham et al. (2014) found that scores from several of these direct measures were predictive of later school readiness, which led these authors to promote their broad use in preschool assessment. Direct measures have important advantages for assessment, especially for internalized constructs (McKown, 2017). For example, McKown noted, "although observers and raters can make educated guesses about children's thinking skills, these skills exist in a child's mind and can't be directly observed" (pp. 168–169). For these and similar constructs (e.g., self-awareness) it can be very difficult for external observers to infer children's skill levels. As such, there is an important role for direct assessment in preschool SEL practice and much research has focused on developing and honing these tools.

Yet, direct assessments also have important limitations. For example, Denham et al. (2014) concluded that traditional use of direct assessment is "time prohibitive" (p. 447) a concern echoed by others (e.g., McKown, 2017). Another assessment modality – rating scales – have distinct advantages in this domain and are considered optimal for assessing behavioral expression of SEL skills (McKown, 2017). Although the CASEL framework has influenced policy and practice, there are surprisingly few sound rating scales of preschool children's SEL skills. For example, the CASEL Assessment Guide (https://measuringsel.casel.org/assessment-guide/) and RAND Assessment Finder (rand.org/education-and-labor/projects/assessments/tool), two major online compendia dedicated to documenting assessments of SEL competencies, collectively list over 50 assessments of children and youth SEL skills, but only four rating scale assessments of preschool children's SEL skills (Table 1). In addition, searches of other comprehensive test resources (e.g., Tests in Print IX; Anderson, Schlueter, Carlson, & Geisinger, 2016) did not yield any additional published preschool SEL rating scales.

Two of the four preschool rating scales are from the SSIS SEL Edition of assessments including the SSIS SEL Rating Form-Teacher and Parent versions (Gresham & Elliott, 2017) and SSIS SEL Screening and Progress Monitoring Scales (Elliott & Gresham, 2017a). The Screening and Progress Monitoring Scales are criterion-referenced performance rubrics designed to be used with the SSIS SEL Classwide Intervention Program (Elliott & Gresham, 2017b). The SSIS SEL Rating Forms are norm-referenced measures, available in both English and Spanish, and allow for a multi-informant (teacher and parent) examination of young children's SEL strengths and areas for improvement. Both the Teacher and Parent versions of SSIS SEL are comprised of 51 items, each of which is aligned with a competency domain in the CASEL framework.

One major advantage of the SSIS SEL Rating Forms is their prominence. The SSIS SEL is a successor to the Social Skills Rating System (SSRS; Gresham & Elliott, 1990), which is arguably the most widely used measure of social competence in preschool children. For example, in a review of over 75 measures of social and emotional development of young (birth through age 5) children, Halle and Darling-Churchill (2016) concluded that only two, the Deveraux Early Childhood Assessment Clinical Form and the SSRS "combine a broad coverage of the subdomains of social and emotional development with strong psychometric properties and ease of administration" (p. 15). It is perhaps due to these advantages that the SSRS and its successor assessments have been so prominent. For example, 14 items from the SSRS have been used for more than a decade by the Department of Education for their widely used Early Childhood Longitudinal Studies (Tourangeau, Nord, Lê, Pollack, & Atkins-Burnett, 2006; Tourangeau et al., 2019) and the SSRS is frequently used in practice (e.g., Wang, Sandall, Davis, & Thomas, 2011). Thus, one advantage of the SSIS SEL forms is their prominence in both research and practice.

Another important advantage of the SSIS family of assessments is their link with intervention programs with evidence of efficacy (e.g., DiPerna, Lei, Cheng, Hart, & Bellinger, 2018) and social validity (Wollersheim Shervey, Sandilos, DiPerna, & Lei, 2017). Specifically, results from SSIS SEL assessments provide a direct actionable link to the SSIS SEL CIP (Elliott & Gresham, 2017), a teacher implemented, evidenced-based program for Pre-K to High School students. The CIP focuses on skills that are assessed by the teachers and parents via the SSIS SEL Rating Forms (e.g., *Listen to others, Follow the rules, Ask for help, Get along with others, Stay calm with others*). These core skills are fundamental to the development of healthy children and recognized as salient during the 3–6 year old development period (Bierman & Motamedi, 2015). As such, the alignment between the SSIS SEL family of assessments and validated interventions present another key advantage.

Despite these important advantages, there are limitations of the SSIS SEL and similar instruments. Most notably, although standard rating scales such as the SSIS SEL are generally more time-efficient than direct measures, they are often still long, and the SSIS SEL is no exception. Specifically, the SSIS SEL is composed of 51 items and takes teachers and parents roughly 15–20 min to complete. This limitation has been noted for assessments similar to the SSIS SEL, including the SSRS (e.g., Halle & Darling-Churchill, 2016). The length of the SSIS SEL is likely especially problematic for applications other than individual decision making. Yet, there are growing calls for broad inclusion of strengths-based preschool assessment to both identify students in need and provide information to inform further assessment and intervention (e.g., Denham et al., 2014; LeBuffe & Shapiro, 2004). Assessment with teacher rating scales such as the SSIS SEL for more than a few students at a time could quickly overburden teachers whose time is already limited. If conducted at the universal level, current measures are likely to be impractical for the vast majority of users – researchers and educators alike. For example, completion of SSIS SEL forms for 10–15 students could easily take 2–4 h of teacher time. Furthermore, long, extensive forms are likely a deterrent for parental completion of forms, limiting the sources from which information regarding student SEL can be gathered. As such, briefer versions of the SSIS SEL appropriate for preschool children might serve to capitalize on the strengths of the measure (CASEL alignment, prominence, alignment with intervention) while addressing its most critical current limitation.

Table 1
Description of published SEL assessments for preschool students.

Assessment (publication date)	Informant(s)	# of items	Competencies assessed	Completion time
Panorama SEL Teacher Rating of Student SEL Competencies (2017)	Teacher	10–62 depending on grade	Classroom Effort, Emotion Regulation, Grit, Growth Mindset, Learning Strategies, Self-Efficacy, Self-Management, Social Perspective Taking, Social Awareness	10–15 min
Six Seconds Perspective Youth Version (2018)	Teacher Family	51	37 competencies including Adaptability, Consequential Thinking, Collaboration, Drive, Emotional Insight, Engage Intrinsic Motivation, Optimism, Focus, Good Health, Imagination, Empathy, Personal Achievement, Problem Solving, Resilience, Self-Awareness, Self-Management	5–10 min
SSIS SEL Edition Screening & Progress Monitoring Scales (2017)	Teacher	8 × 5-level performance rubrics	Self-Awareness, Self-Management, Social Awareness, Relationship Skills, Responsible Decision Making; Motivation to Learn, Early Reading, Early Mathematics	2 mins per student; 35–40 min per class
SSIS SEL Edition Rating Forms (2017)	Teacher Parent	51 SEL + 7 Academic Competence[a]	Self-Awareness, Self-Management, Social Awareness, Relationship Skills, Responsible Decision Making; Academic Competence[a]	10–15 min

Note. All assessments can be used with preschool – 12th grade students.
[a] Academic competence items of SSIS SEL Edition Rating Form are only included in the teacher form of the measure.

1.3. Item response theory and the development of efficient assessments

In recent years, the need for efficient assessments of children's social, emotional, and behavioral competencies has been raised in other fields and with older student age groups (e.g., Anthony, DiPerna, & Lei, 2016; Anthony et al., 2020; Gresham et al., 2010). Indeed, several investigations have utilized advanced psychometric procedures to identify sets of items from full-length forms for this very purpose (e.g., Anthony & DiPerna, 2017, 2018; Moulton, von der Embse, Kilgus, & Drymond, 2019). Such investigations have increasingly relied on item response theory (IRT) to achieve these goals. Put briefly, IRT is a psychometric approach centered on modeling the relationship between the latent construct being measured (e.g., self-management) and one or more features of each item (e.g., the item's difficulty). Scale development grounded in IRT has several advantages relative to traditional approaches grounded in classical test theory (CTT) including sample free estimation of item parameters (provided adequate model fit and satisfaction of model assumptions), the ability to test the plausibility of alternate measurement models and the production of visual output demonstrating item function along the latent trait continuum (Anthony et al., 2016). Perhaps most relevant for the process of improving measurement efficiency, however, is the production of item and test information functions.

In IRT, the term *information* refers to score precision and is akin to reliability in CTT. Unlike CTT, however, information is not a static feature of tests in IRT, but rather varies across the latent construct being assessed (e.g., self-management). This feature of IRT allows test developers to identify which items contribute the most to precision at which point on the latent construct scale. For example, using information functions, test developers can select items with high information in the "at risk" trait range for the purposes of educational screening. Such use of IRT allows for the selection of items which are most efficient and thereby drastically shorten test length while retaining much of the precision of their parent forms, especially in the most important ranges of targeted latent traits (e.g., lower ends of trait continua characterizing students at risk for persistent difficulty). This feature of IRT is especially helpful for the development of efficient forms of longer measures (Anthony et al., 2016).[1]

There are also important advantages of IRT for evaluating and promoting the fair and unbiased use of assessments. Specifically, IRT enables a fine-grained evaluation of Differential Item Functioning (DIF; Meade, 2010; Tay, Meade, & Cao, 2015). Succinctly, DIF occurs when item functions (e.g., difficulties) differ across various demographic groups when holding latent trait level constant. Thus, mere mean differences on item scores do not necessarily indicate DIF. For example, an item in which girls who had the same overall self-management skills as boys scored substantially lower than boys not attributable to chance error would likely be flagged for sex-based DIF. By itself, DIF does not necessarily indicate bias, but it always warrants careful attention and consideration. Fortunately, IRT enables fine grained evaluation of DIF including the consideration of effect sizes for the magnitude of DIF and the production of indices and plots to determine the construct levels likely to be influenced by retention of DIF items.[2]

1.4. Purpose and research questions

Although brief versions of K-12 SEL rating scales recently have been developed (Anthony et al., 2020), no similar measures have been developed for use with preschoolers. Thus, in the remainder of this article, we report on the application of IRT to refine the SSIS SEL Rating Forms into the more efficient SSIS SEL Brief Scales – Teacher Preschool Form (SSIS SEL*b*–TP) and SSIS SEL Brief Scales – Parent Preschool Form (SSIS SEL*b*-PP). In creating the SSIS SEL*b*-TP and SSIS SEL*b*-PP, we aimed to: (a) significantly reduce the length of the full length SSIS SEL assessments, (b) retain appropriate content coverage of SSIS SEL constructs, (c) produce scales yielding scores with sufficient reliability for low stakes decisions, (d) pro-

[1] For further reference on IRT in general, we recommend De Ayala (2013), or Embretson and Reise (2000). For further reference on the use of IRT to develop efficient versions of longer measures, we recommend Anthony et al. (2016).
[2] For further reference on DIF, we recommend De Ayala (2013), Meade (2010), and Tay et al. (2015).

Table 2
Demographic characteristics of participants (percentages).

students characteristic	SSIS SEL*b*-TP (*N* = 341)	SSIS SEL*b*-PP (*N* = 723)	Current U.S. preschool population
Female	47	48	49
Race			
White	62	71	49
Black	11	10	14
Hispanic	19	13	26
Other	8	6	10
Region			
Northeast	11	19	–
Midwest	33	29	–
South	49	39	–
West	8	13	–
Parent's education level			
Grade 11 or less	11	8	9
Grade 12 or GED	22	21	21
1–3 years of college	31	35	29
4+ years of college	37	36	42
Educational status			
General education	88	79	–
Special education	12	21	–

Note. Some percentages do not sum to 100 due to rounding. SSIS SEL*b*-TP = SSIS SEL Brief Scales – Teacher Preschool Form; SSIS SEL*b*-PP = SSIS SEL Brief Scales – Parent Preschool Form. Preschool population estimates from the 2018 Digest of Educational Statistics (Snyder, de Brey, & Dillow, 2019), which does not report data for Region or Educational Status.

duce scales yielding scores with initial evidence of validity, and (e) generate scales with evidence indicating a lack of item and test bias.

2. Method

2.1. Participants

Participants were teachers and parents of all 3- to 5-year-old preschool students collected for the SSIS standardization sample. Although the standardization sample included 200 cases rated by teachers and 400 cases rated by parents, additional cases were collected such that cases could be carefully selected to match census population estimates (i.e., in the initial standardization data collection, more cases than were needed for standardization were collected). Because our planned IRT analyses function best with large sample sizes (De Ayala, 2013), we utilized all available cases (i.e., those included in the standardization sample as well as additional cases collected at the same time but not ultimately chosen for the standardization sample). These additional cases rendered the total number of cases 341 for the SSIS SEL – Teacher (SSIS SEL-T; Gresham & Elliott, 2017) and 723 for the SSIS SEL – Parent (SSIS SEL-P; Gresham & Elliott, 2017). These cases were diverse across race/ethnicity, sex, and parent education level. Full demographic data for these cases are reported in Table 2.

2.2. Measures

SSIS Social Emotional Learning Edition Rating Form – Teacher. The SSIS SEL-T (Gresham & Elliott, 2017) is a nationally normed behavior rating scales of SEL for students ages 3–18. The SSIS SEL-T includes 58 items rated on a 4-point Likert scale from 0 (*Never*) to 3 (*Almost Always*); 51 items measure SEL skills. The remaining 7 items focus on academic competence and are not completed at the preschool level. With regard to reliability, coefficient α's for students ages 3–5 ranged from .77 to .97, with a median value of .90 across the five SSIS SEL scales and the SEL composite. Furthermore, 2-month stability coefficients for a sample of students were in the low .80s; mean scores between administrations were very similar, with most effect sizes under .10. Although multiple sources of evidence are reported in the technical manual to support the validity of SSIS SEL-T scores for students ages 3–18, evidence for 3–5 year olds was not examined separately. Finally, confirmatory factor analyses (CFA) also provided support for the internal structure of the SSIS SEL-T yielding a six-factor model, five of which represented the CASEL SEL competencies and a sixth factor representing Academic Competence (Gresham, Elliott, Byrd, Wilson, & Cassidy, 2018; Gresham et al., 2018b).

SSIS Social Emotional Learning Edition Rating Form – Parent. The SSIS SEL-P (Gresham & Elliott, 2017) is a nationally normed behavior rating scales of SEL for students ages 3–18. The SSIS SEL-P includes 51 items rated on a 4-point Likert scale from 0 (*Never*) to 3 (*Almost Always*); these are the same 51 items measured on the SSIS SEL teacher version and are rated on the same Likert scale. With regard to reliability, coefficient α's for students ages 3–5 ranged from .75 to .96, with a median value of .88 across the five SSIS SEL-P scales and the SEL composite. Furthermore, 2-month stability coefficients for SEL subscales were in the upper .70s and low .80s; mean scores between administrations were very similar, with most effect sizes under .10, indicating very stable performance across the testing interval. Substantial evidence is reported to support the validity of SSIS SEL-P scores for samples of students ages 3–18, but subsample evidence for 3–5 year olds is not provided separately. Finally, CFAs also provided support of the internal structure of the SSIS SEL-P yielding a five-factor model consistent with the CASEL SEL competencies (Gresham, Elliott, Metallo, et al., 2018).

Social Skills Rating System Teacher and Parent Forms. The preschool version of the SSRS (Gresham & Elliott, 1990) were used as validity measures in this study. The Social Skills Scale of the preschool SSRS-Teacher (SSRS-T) is comprised of three subscales: Cooperation, Assertion, and Self-Control. The SSRS-Parent (SSRS-P) includes a fourth subscale of Responsibility. Using a 3-point response scale, parents and teachers rate each social skills item based on the frequency of the behavior. Response options include *Never, Sometimes,* or *Very Often.* The Problem Behaviors Scale consists of Externalizing and Internalizing behavior subscales. The Problem Behaviors Scale was intended to function as a screener, focusing on only 10 problem behavior items. Parents and teachers rate the frequency of each behavior as *Never, Sometimes,* or *Very Often.* A sample (N = 200) of preschool children primarily from two large metropolitan areas, one in the Southeastern and the other in the Midwestern United States, was used to evaluate the psychometric characteristics of the preschool SSRS-T and SSRS-P.

Evidence for the internal structure of the preschool SSRS was established when the factor analysis of the scales conformed to that found with a much larger, nationally representative elementary sample (Gresham & Elliott, 1990). Subsequent scale and subscale

inter-correlations and item-total correlations also provided evidence the preschool scales psychometrically were functioning very similarly to the elementary version of scales based on a nearly identical pool of items (e.g., Frey, Elliott, & Gresham, 2011). Specifically, the SSRS-T Preschool version has high internal consistency for the total score (α = .93) and test-retest reliability of .85. The SSRS-P version has high internal consistency for the total score (α = .90) and test–retest reliability of .87. Gresham and Elliott (1990) reported that each of the three prosocial skills scales on the SSRS-T and SSRS-P correlated at a moderate negative level with the Walker Problem Behavior Identification Checklist. Gresham, Elliott, and Black (1987) showed that the SSRS-T ratings on all factors were free from rater racial bias and sex bias.

2.3. Procedures

Data used in the current study were collected as part of the original SSIS Rating Scale standardization. Pearson Assessment field staff recruited site coordinators in 21 schools across 15 states, who in turn, recruited participants to fit demographic targets. These site coordinators and their preschools distributed and collected the rating scales from September 2006 to October 2007. The final standardization sample was selected using a stratified random sampling approach from the larger respondent sample to fit 2006 U.S. Census demographics of age, sex, race/ethnicity, and educational status.

2.4. Data analyses

Data analyses proceeded in several phases, including checking IRT assumptions, item analysis, initial reliability analyses, and initial validity analyses. We first evaluated standard IRT assumptions, including the assumption of unidimensionality and local independence. First, standard IRT analyses assume that the targeted latent construct accounts for the majority of variance in items scores (i.e., that the item set is essentially unidimensional; Anthony et al., 2016). We evaluated unidimensionality by SSIS SEL scale using exploratory factor analysis (EFA) conducted in MPlus (Muthen & Muthen, 2017). In line with recommendations for item level analyses for polytomous data with 4 or fewer categories (Rhemtulla, Brosseau-Liard, & Savalei, 2012), we treated item level data as categorical. We considered scales to be essentially unidimensional if the ratio of the first to second eigenvalues exceeded 4 (Reeve, Hays, Chang, & Perfetto, 2007). In cases in which this threshold was not met, the lowest loading item was eliminated from consideration until all scales met this assumption. Next, IRT assumes that, controlling for the latent construct, items are not overly related. Violations of local independence could occur if, for example, items were redundant among other reasons (e.g., items such as "says please" and "says thank you"; Anthony et al., 2016). With regard to the assumption of local independence, we utilized standardized local dependence χ^2 indices produced by IRTPro (Cai, Thissen, & du Toit, 2019). When item pairs evidence local dependence (evaluated with a threshold of 10 as recommended by Cai et al., 2019), one of the items was excluded from further consideration such that the SSIS SEL*b* forms had no items with evidence of local dependence.

Once these assumptions had been checked, we conducted IRT analyses. In line with similar investigations (e.g., Anthony et al., 2016; Moulton et al., 2019), we employed the Graded Response Model (GRM; Samejima, 1969) using IRTPro version 4 (Cai et al., 2019). We evaluated model fit with primary reference to RMSEA, with values less than .10 indicating adequate fit to the GRM (MacCallum, Browne, & Sugawara, 1996). We then utilized the item information functions (IIFs) resulting from GRM analyses as the primary psychometric indicator of item adequacy when completing item analysis. Our major goal in this process was to select items that resulted in limited information loss overall and kept information above the .80 reliability threshold recommended for individual screening decisions (Salvia, Ysseldyke, & Witmer, 2016). This reliability threshold corresponded with an information level of 5 based on a formula demonstrated by Petrillo, Cano, McLeod, and Coon (2015) that converts IRT information into a standard reliability metric. In anticipation of probable use of the SSIS SEL*b* with students experiencing some difficulty, we specifically focused on the "at risk" range, which we defined as from .5 to 1.5 standard deviations below the mean (i.e., −0.5 to −1.5 on the θ scale).

In addition to item information, we considered several other indications of item quality when selecting items for the SSIS SEL*b* forms. First, we considered item content to attempt to ensure a close alignment between the SSIS SEL*b* scales and corresponding CASEL domains. Furthermore, we evaluated items relative to the preschool developmental behavioral expressions that would reflect CASEL competencies. Indeed, as a result of consideration of the content of the SSIS SEL Self Awareness scales relative to the developmental level of preschool children, we opted to *not* select items for a brief Self-Awareness scale due to questions regarding the developmental appropriateness of the items on the SSIS SEL for preschool-age children. This challenge is not unique to the SSIS SEL as self-awareness and other similar domains are very difficult for third parties to assess in young children.

For example, in their review of the state of social and emotional skill measurement, Humphrey et al. (2011) identified 12 measures for review, none of which included a domain labeled "self-awareness." They argued that there are difficulties measuring this construct with assessments targeting children's typical behavior (as opposed to their optimal behavior), such as most rating scales. Furthermore, these authors argued that constructs with an internal locus such as self-awareness are most appropriately assessed by individuals themselves, but that very young students are unlikely to be able to accurately self-report competence in these domains. Thus, it is not surprising that item content for the SSIS SEL precluded development of a self-awareness scale. However, this represents an important goal for future development. Finally, we attempted to maximize the item alignment across the SSIS SEL*b*-TP and SSIS SEL*b*-PP to support the usage of the measure with multiple informants.

Beyond content, we also completed differential item functioning (DIF) analyses for sex (girls vs. boys) and race/ethnicity (White vs. Nonwhite) to facilitate item selection. Specifically, we utilized a two-step purification procedure (Tay et al., 2015) to identify items that had statistical evidence of DIF. When items were flagged as exhibiting DIF, we also completed further analyses to evaluate the magnitude of DIF. We utilized the Expected Score Standardized Difference (ESSD) index, which indicates the overall standardized differences between expected scores for focal group participants when calculated using the focal or reference group parameters (Tay et al., 2015). These statistics were calculated using the parameters generated by IRTPro and the Visual DF excel macro (Meade, 2010). These statistics can be interpreted according to Cohen's criteria (i.e., 0.2 = small; 0.5 = medium; 0.8 = large; 1988). Thus, in cases in which content considerations led us to retain items with statistical DIF, we ensured that it was minimal or balanced by other items. Finally, when we did retain items with evidence of some DIF, we evaluated the overall impact of these retentions at the scale level by calculating and plotting the expected scores of relevant group comparisons in cases in which DIF arose (i.e., the sum score that would be expected on a SSIS SEL*b* scale across the continuum of the targeted construct based on estimated item parameters, which will differ in cases of DIF). Such evaluation helped us further evaluate the scope and potential negative consequences of DIF.

Using the aforementioned indices and considerations, each of the authors independently selected four to five items that they con-

Table 3
Summary of item selection criteria on the SSIS SEL Brief Scales – Teacher Preschool and example items.

Domain/item	Overlaps with SSIS SEL*b*-PP item	Local dependence	DIF (ESSD)	
			Race	Sex
Self-management				
7. Completes tasks	√	*ns*	*ns*	*ns*
24. Pays attention	√	*ns*	*ns*	*ns*
33. Stays calm	√	*ns*	*ns*	*ns*
35. Follows rules		*ns*	*ns*	*ns*
Social awareness				
3. Comforts others	√	*ns*	*ns*	*ns*
12. Feels bad when others sad		*ns*	*ns*	*ns*
36. Shows concern	√	*ns*	*ns*	*ns*
46. Stands up for others	√	*ns*	*ns*	*ns*
Relationship skills				
1. Makes friends	√	*ns*	*ns*	*ns*
18. Interacts well	√	*ns*	*ns*	*ns*
28. Invites others	√	*ns*	*ns*	*ns*
37. Starts conversations	√	*ns*	*ns*	*ns*
Responsible decision-making				
2. Takes responsibility	√	*ns*	*ns*	*ns*
9. Is well-behaved	√	*ns*	√(−0.02)	*ns*
21. Acts responsibly.		*ns*	*ns*	*ns*
25. Takes care when using things	√	*ns*	*ns*	*ns*

Note. SSIS SEL*b*-PP = SSIS SEL Brief Scales – Parent Preschool Form. DIF = Differential Item Functioning. ESSD = Expected Score Standardized Difference. *ns* = presence of item issue not indicated according to a priori criteria. All item numbering derived from Social Skills Improvement System – Social and Emotional Learning Edition. Item stems are slightly abbreviated and modified so as not to violate copyright but reflect the content of each item.

sidered good candidates for each SSIS SEL*b* scale. Then, through a cyclical process of discussion and consideration of content and psychometric quality indicators, four items were identified for each SSIS SEL*b* scale. Once these sets of items were identified, we completed initial reliability and validity analyses on them. First, we examined test information functions (TIFs) for each identified scale. Next, we computed Cronbach's α on each scale as another overall metric of reliability. We also utilized the test–retest and interrater reliability standardization samples to compute these statistics for identified SSIS SEL*b* items. Samples were 48 and 49 children for Teacher and Parent test-retest reliability analyses, respectively, and 28 for Parent form interrater reliability analyses (i.e., the child's two parents). In addition to computing these initial reliability analyses for SSIS SEL*b* items, we also computed these statistics with the full SSIS SEL Teacher and Parent forms for comparative purposes.

Next, we conducted initial convergent and discriminant validity analyses on identified SSIS SEL*b* forms. First, we computed interscale correlations for all SSIS SEL*b* scales and composites. Given the interrelations between the SEL constructs represented on the SSIS SEL*b*, we anticipated moderate to strong intercorrelations (i.e., $r > .30$) between these scales. We then correlated SSIS SEL*b* scales with scale scores from the SSRS, including the Cooperation, Assertion, Self-Control, Externalizing, and Internalizing scales for both teacher and parent forms of the SSRS and the addition of the Responsibility scale for the parent form of the SSRS. We anticipated moderate to strong (i.e., $r > .30$) positive relations between SSIS SEL*b* scales and SSRS scales indicating prosocial behaviors (i.e., the SSRS Cooperation, Assertion, Responsibility, and Self-Control scales) and moderate to strong negative relations (i.e., $r < -.30$) between SSIS SEL*b* scales and SSRS scales indicating problem behaviors (i.e., the SSRS Externalizing and Internalizing scales).

Additionally, we calculated validity coefficients with the SSIS SEL Teacher and Parent forms and compared these coefficients to those calculated with the SSIS SEL*b*-TP and SSIS SEL*b*-PP. We determined whether differences between validity coefficients calculated with the SSIS SEL and SSIS SEL*b* were statistically significant using an online calculator developed by Lee and Preacher (2013) that implements Steiger's (1980) formula for comparing two dependent correlations. Finally, we computed the correlations between corresponding scales on the SSIS SEL*b*-TP and SSIS SEL*b*-PP with the available sample ($N = 288$). Given small to moderate positive correlations often are observed between informants (De Los Reyes et al., 2016), we anticipated that these correlations would be small to moderate and positive.

3. Results

3.1. Item analysis, content analysis, and DIF analysis

First, prior to conducting IRT analyses, we evaluated standard IRT assumptions. For unidmensionality analyses, ratios of first to second eigenvalues ranged from 3.52 to 8.72 for the SSIS SEL-T and from 4.35 to 7.89 for the SSIS SEL-P. One scale (the Responsible Decision Making Scale) on the SSIS SEL-T required eliminating the item with the lowest loading to achieve essential unidimensionality. This item (*Stand up for herself/himself when treated unfairly*) potentially had a low loading due to its specific focus on fairness relative to other items that focus more on day-to-day responsibility (e.g., *Takes responsibility for her/his own actions).* Once this item was eliminated, all ratios of first to second eigenvalues met a priori criteria and ranged from 5.22 to 8.72 for the SSIS SEL-T.

Next, we computed GRM analyses for all scales and evaluated local dependence. Across SSIS SEL scales, the percentage of item pairs with standardized χ^2 values exceeding 10 ranged from 0% to 4% for the SSIS SEL-T and from 4% to 17% for the SSIS SEL-P. These violations were addressed during item selection. Finally, we conducted DIF analyses for the SSIS SEL-T and SSIS SEL-P. For the SSIS SEL-T, the number of DIF violations was small as there were no items with evidence of sex-based DIF across scales and only 4 items with evidence of race-based DIF. The magnitude of DIF for these items also was generally small with ESSD values ranging from −0.08 to 0.37 (Median = 0.06). For the SSIS SEL-P, there were more items with DIF for both sex and race. Specifically, the number of items with evidence for sex-based DIF ranged from 0 to 2 (Median = 1) across scales and the number of items with evidence for race-based DIF ranged from 1 to 8 (Median = 4.5) across scales. Despite the higher number of items with evidence of DIF, the magnitude of DIF on most potentially problematic items was small with ESSD values ranging from −0.11 to 0.18 (Median = 0.05) for sex-based DIF items and from −0.41 to 0.36 (Median = −0.08) for race-based DIF items. As with local dependence, these indices were considered during item selection.

Table 4
Summary of item selection criteria on the SSIS SEL Brief Scales – Parent Preschool and example items.

Domain/item	Overlaps with SSIS SEL*b*-TP item	Local dependence	DIF (ESSD)	
			Race	Sex
Self-management				
1. Pays attention	√	*ns*	*ns*	*ns*
26. Completes tasks	√	*ns*	√(−0.08)	*ns*
44. Stays calm	√	*ns*	√(−0.04)	*ns*
50. Follows rules in game		*ns*	*Ns*	*ns*
Social awareness				
15. Shows concern	√	*ns*	√(−0.16)	*ns*
34. Comforts others	√	*ns*	√(−0.08)	*ns*
36. Stands up for others	√	*ns*	ns	*ns*
38. Understands how others feel		*ns*	*ns*	*ns*
Relationship skills				
8. Interacts well	√	*ns*	*ns*	*ns*
16. Makes friends	√	*ns*	*ns*	*ns*
30. Starts conversations	√	*ns*	√(−0.03)	*ns*
37. Invites others	√	*ns*	*ns*	*ns*
Responsible decision-making				
6. Takes care when using things	√	*ns*	*ns*	*ns*
9. Follows rules		*ns*	*ns*	*ns*
13. Is well-behaved	√	*ns*	*ns*	*ns*
29. Takes responsibility	√	*ns*	*ns*	*ns*

Note. SSIS SEL*b*-TP = SSIS SEL Brief Scales – Teacher Preschool Form. DIF = Differential Item Functioning. ESSD = Expected Score Standardized Difference. *ns* = presence of item issue not indicated according to a priori criteria. All item numbering derived from Social Skills Improvement System – Social and Emotional Learning Edition. Item stems are slightly abbreviated and modified so as not to violate copyright but reflect the content of each item.

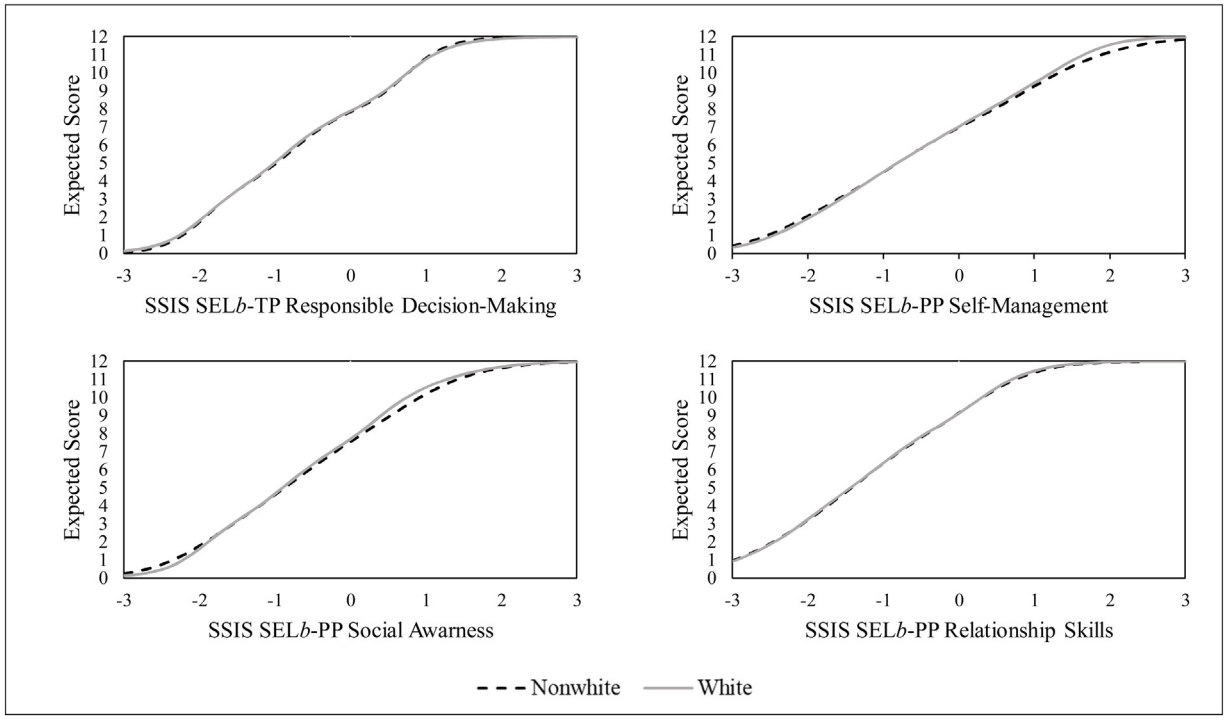

Fig. 1. Expected scores across focal (Nonwhite) and reference (White) groups for SSIS SEL*b*-Teacher Preschool (SSIS SEL*b*-TP) and SSIS SEL*b*-Parent Preschool (SSIS SEL*b*-PP) forms with evidence of race-based differential item functioning.

As a result of item and content analyses, four items per scale were selected for the SSIS SEL*b*-TP and SSIS SEL*b*-PP. All psychometric quality indicators can be found in Tables 3 and 4. On the SSIS SEL*b*-TP, there were no instances of LD or sex-based DIF. Due to content consideration, one item on the SSIS SEL*b*-TP with evidence of race-based DIF was retained, but the magnitude of DIF on this item was very small (ESSD = −0.02). Plotting the expected scores across groups (Fig. 1) indicated that there is no trait level at which this level of DIF appears to make appreciable differences in expected SSIS SEL*b*-TP Responsible Decision Making scores. Next, as with the SSIS SEL*b*-TP, there were no instances of LD or sex-based DIF on the SSIS SEL*b*-PP. However, due to content considerations and the larger number of items with indications of race-based DIF, the final SSIS SEL*b*-PP had five items with evidence of race-based DIF. Despite this finding, the magnitude of DIF was small in all cases (ESSD values ranged from −0.16 to −0.03; median = −0.08).

The effects of retention of these DIF items on the relevant SSIS SEL*b*-PP scale scores was evaluated further by plotting expected scores across groups (Fig. 1). As with the SSIS SEL*b*-TP, there were few levels at which DIF of retained items would lead to any SSIS SEL*b*-PP score differences between White and Nonwhite students. Yet, it appears that should differences due to DIF arise, they would

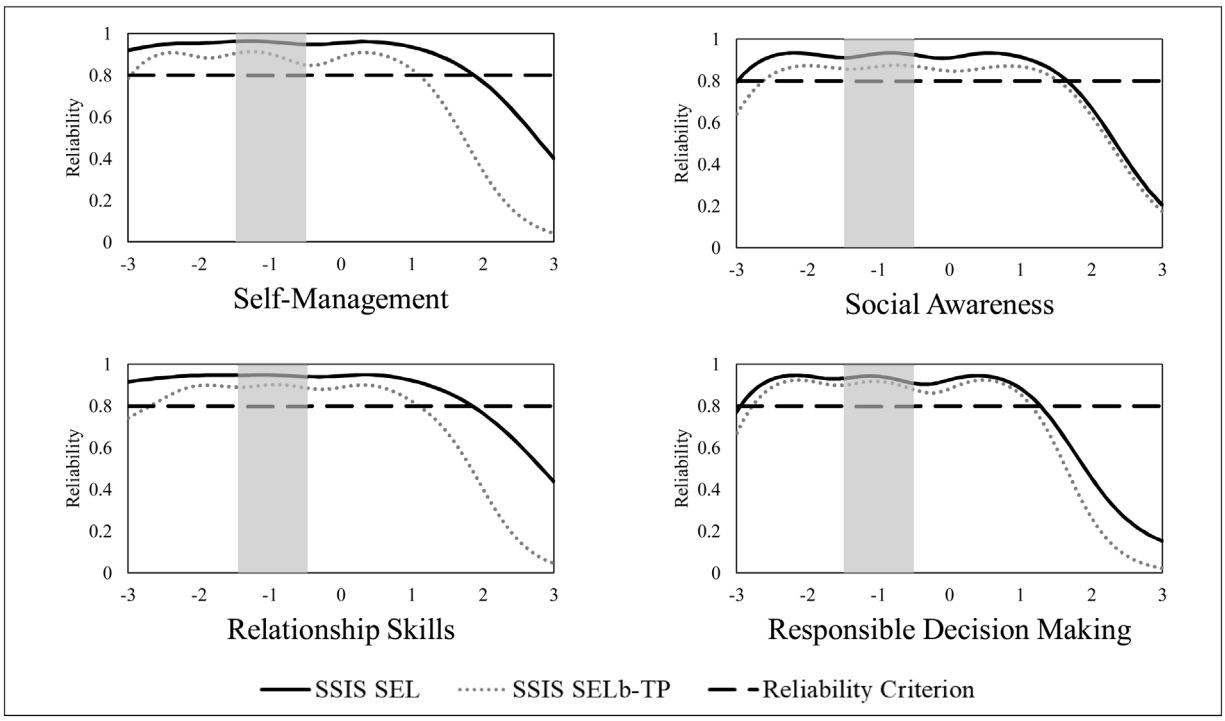

Fig. 2. Test Information Functions for Social Skills Improvement System – Social Emotional Learning Edition (SSIS SEL) and the SSIS SEL Brief Scales – Teacher Preschool Form (SSIS SEL*b*-TP). *Note.* Reliability on *y*-axis converted from total information with the following formula: 1 – (1/information) as recommended by Petrillo et al. (2015). Shaded region represents "at risk" range.

favor White students in the high range (about 1.5–2.5 on the θ scale) of the SSIS SEL*b*-PP Self-Management scale and in the average-high range (about 0.5–1.25 on the θ scale) of the SSIS SEL*b*-PP Social Awareness scale. Furthermore, DIF could lead to slight differences favoring Nonwhite students in the low ranges (about −3 to −2.5 on the θ scale) of both of these scales. There was no area in which retained DIF items appear to lead to any appreciable difference between White and Nonwhite students on the SSIS SEL*b*-PP Relationship Skills scale. Finally, in addition to the technical properties of these items, 81% of included items on the SSIS SEL*b*-TP and SSIS SEL*b*-PP were overlapping in content (i.e., were equivalent items slightly adapted to either teachers or parents).

3.2. Reliability analyses

Next, we conducted initial reliability analyses for SSIS SEL*b* scales. The primary sources of reliability evidence were the TIFs showing reliability across the spectrum of the SEL constructs targeted by the SSIS SEL*b*. As shown in Figs. 2 and 3, compared with their full length counterparts, the SSIS SEL*b* scales retained fairly high levels of precision across a broad range of the latent trait spectrum. In general, reliability levels exceeded .80 for the SSIS SEL*b* scales from roughly −3 to 1.5 on the latent trait scale. This was similar to full SSIS SEL counterparts, which generally had reliability levels exceeding .80 from −3 to roughly 2.5 on the latent trait scale. Furthermore, at no point did reliability fall below the .80 a priori criterion in the at risk range for any SSIS SEL*b* scale. With regard to traditional reliability indices, α was .92 for the SEL composite of both the SSIS SEL*b*-TP and the SSIS SEL*b*-PP. Across SSIS SEL*b*-TP and SSIS SEL*b*-PP scales, α coefficients ranged from .76 to .83 (Median = .80). Next, test-retest reliability coefficients for the SEL composite were .93 and .81 for the SSIS SEL*b*-TP and SSIS SEL*b*-PP scales respectively. Across SSIS SEL*b*-TP and SSIS SEL*b*-PP scales, test-retest reliability coefficients ranged from .59 to .92 (Median = .86). Finally, the interrater reliability coefficient for the SSIS SEL*b*-PP composite was .67, and interrater reliability coefficients ranged from .51 to .71 (Median = .54) across SSIS SEL*b*-PP scales. These indices, along with comparison indices from the SSIS, are reported in Table 5. In general, scores from SSIS SEL*b* scales were slightly less reliable than SSIS SEL counterparts. Across scales these differences ranged from −.12 to −.03 (median = −.08) for α coefficients, from −.21 to .04 (median = 0) for test-retest coefficients, and from -.15 to .02 (median = −.05) for parent form interrater reliability coefficients. Differences were also slight across SEL composites, with reliability coefficient differences ranging from −.04 to .01 (median = .03) across all coefficient types and forms.

3.3. Validity analyses

Finally, we conducted initial validity analyses for SSIS SEL*b* scales, which consisted of scale intercorrelations and validity correlations with the SSRS (Table 6). All intercorrelations for SSIS SEL*b* scales were strong, ranging from .57 to .86 (median = .62) for the SSIS SEL*b*-TP and from .53 to .78 (median = .59) for the SSIS SEL*b*-PP. Furthermore, correlations between the SSIS SEL*b* composite scores and the SSIS SEL*b* scales were also strong ranging from .84 to .89 (median = .87) for the SSIS SEL*b*-TP and from .81 to .87 (median = .84) for the SSIS SEL*b*-PP. Differences between these interscale and composite-scale correlations and corresponding correlations calculated with the SSIS SEL ranged from −.22 to −.03 (median = −.09; negative values indicate that SSIS SEL*b* intercorrelations were weaker than SSIS SEL intercorrelations) for the SSIS SEL*b*-TP and from −.22 to −.12 (median = −.02) for the SSIS SEL*b*-PP.

With regard to validity coefficients, correlations generally were in line with expectations as well. Correlations between SSIS SEL*b* scales/composite and SSRS social skills scales were moderately to strongly positive ranging from .52 to .78 (median = .67) for the SSIS SEL*b*-TP and from .31 to .69 (median = .55) for the SSIS SEL*b*-PP. Correlations also were in line with expectations for the SSRS Externalizing scale, but lower than expected in magnitude for the SSRS

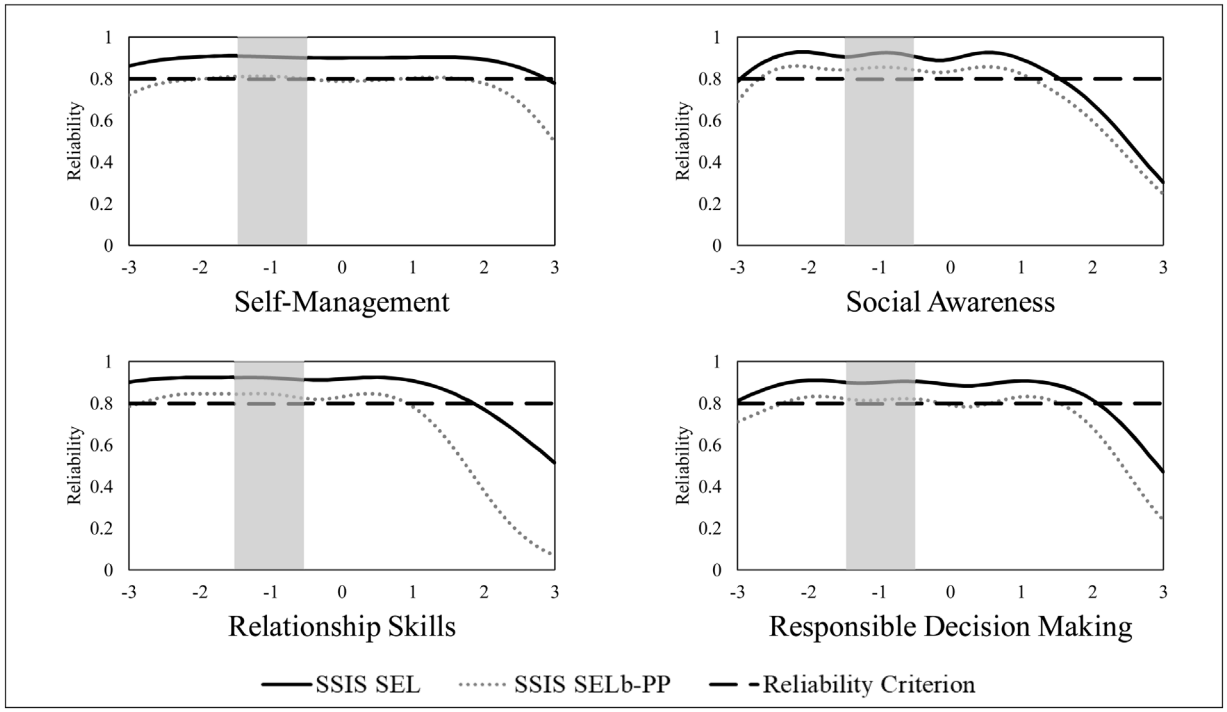

Fig. 3. Test Information Functions for Social Skills Improvement System – Social Emotional Learning Edition (SSIS SEL) and the SSIS SEL Brief Scales – Parent Preschool Form (SSIS SEL*b*-PP). *Note.* Reliability on *y*-axis converted from total information with the following formula: 1 – (1/information) as recommended by Petrillo et al. (2015). Shaded region represents "at risk" range.

Table 5
SSIS SEL Brief Scales – Preschool Forms and SSIS SEL preschool reliability statistics for teacher and parent forms.

Scale	Cronbach's α				Test–retest				Parent inter-rater(n = 28)	
	Teacher (n = 341)		Parent (n = 723)		Teacher (n = 48)		Parent (n = 49)			
	SEL*b*	SEL	SEL*b*	SEL	SEL*b*	SEL	SEL*b*	SEL	SEL*b*	SEL
Self-management	.78	.87	.79	.82	.92	.92	.67	.69	.51	.66
Social awareness	.76	.88	.82	.88	.85	.84	.70	.78	.53	.51
Relationship skills	.81	.89	.83	.90	.90	.90	.86	.84	.54	.62
Responsible decision making	.81	.86	.79	.87	.91	.87	.59	.80	.69	.71
SEL composite	.92	.96	.92	.96	.93	.92	.81	.84	.67	.67

Note. SEL*b* = SSIS SEL Brief Scales – Preschool Forms; SEL = Social Skills Improvement System Social Emotional Learning Edition; Interrater statistic is for two parents of each included child.

Table 6
SSIS SEL Brief Scales – Preschool Forms and SSIS SEL validity coefficients.

	Self-management		Social awareness		Relationship skills		Responsible decision making		SEL composite	
	SEL*b*	SEL	SEL*b*	SEL	SEL*b*	SEL	SEL*b*	SEL	SEL*b*	SEL
SSRS Teacher (n = 98)										
Cooperation	.67	.65	.52	.53	.69	.74	.56	.63	.72	.71
Assertion	**.57**	**.65**	.69	.67	.77	.83	.59	.74	.78	.79
Self-Control	**.70**	**.77**	.65	.68	.56	.71	.61	.68	**.74**	**.78**
Externalizing	−.70	−.67	−.48	−.50	−**.33****	−**.49**	−**.66**	−**.57**	−.63	−.61
Internalizing	−.33**	−.35	−.23*	−.24*	−.53	−.52	−.21*	−.29**	−.39	−.40
SSRS Parent (n = 118)										
Cooperation	.59	.60	.57	.57	**.35**	**.50**	.67	.64	.67	.65
Assertion	**.31****	.36	.43	.40	.56	.55	.31**	.34	.50	.48
Responsibility	.50	.51	.55	.53	**.39**	**.50**	.54	.57	.61	.59
Self-Control	.69	.73	.47	.47	**.32**	**.51**	**.69**	**.62**	.67	.67
Externalizing	−**.60**	−.61	−.39	−.42	−**.31****	−**.48**	−.59	−.54	−.58	−.59
Internalizing	−.24**	−.23*	−.08$^{n.s.}$	−.13 $^{n.s.}$	−.31**	−.32	−.15 $^{n.s.}$	−.15 $^{n.s.}$	−.23*	−.25**
Parent–Teacher Correlations (n = 288)	.35	.34	.36	.39	.48	.49	.47	.43	.47	.49

Note. SEL*b* = SSIS SEL Brief Scales – Preschool Forms; SEL = Social Skills Improvement System Social Emotional Learning Edition; SSRS = Social Skills Rating System. Bolded coefficients indicate statistically significant differences between validity coefficients. Unless otherwise noted, all correlations are statistically significant (p < .001).
$^{*}p$ < .05.
$^{**}p$ < .01.
$^{n.s.}$ p > .05.

Internalizing scale. Specifically, across parent and teacher scales and composites, SSIS SEL*b* validity coefficients with the SSRS Externalizing scale ranged from −.70 to −.31 (Median = −.59). In contrast, for the SSRS Internalizing scale, these coefficients ranged from −.53 to −.08 (Median = −.24). Validity coefficients generally were similar when calculated with the SSIS SEL and the SSIS SEL*b*. Specifically, validity correlation differences ranged from −.15 to .16 (median = -.02) for the SSIS SEL*b*-TP and from −.19 to .17 (median = .01) for the SSIS SEL*b*-PP. Thus, in general, validity coefficients largely were in line with expectations and similar when calculated with the SSIS SEL*b* and the SSIS SEL.

4. Discussion

The purpose of this study was to develop and initially validate a brief preschool version of the SSIS SEL, one of the few measures of preschool children's SEL competencies aligned with the influential CASEL framework for many states' SEL standards. Our major goals included developing efficient and fair teacher and parent measures that retain evidence of score reliability and validity. Overall, these goals were met, although there were some limitations. The resulting SSIS SEL*b* Scales hold promise to capitalize on teachers' and parents' observations and extend the assessment options for preschool children's SEL skills across the country.

First, reliability estimates were moderately strong and sufficient to support low-stakes individual or group decisions (Salvia, Ysseldyke, & Witmer, 2017). This conclusion is supported both by our IRT analyses as well as more traditional reliability metrics. As shown in Figs. 2 and 3, our analyses indicate that the SSIS SEL*b* would likely function best for students with difficulties or moderate strengths in the assessed SEL skill domains. The SSIS SEL*b* did not demonstrate high precision at higher levels of SEL skills and would likely not function as well when assessment at these levels is the priority. This finding is similar to other rating scale research employing IRT for older children (e.g., Anthony et al., 2016), which found levels of score precision similarly tapered at higher levels of targeted traits. Indeed, this pattern has been found when IRT analyses are used for many psychological traits (Reise & Waller, 2009). It is possible that new approaches to rating scale item design are necessary to better assess significant student strengths in these domains. Such development would be particularly important for the SEL domain, which is strongly focused on student strengths. Regardless of whether such developments occur, these findings illustrate the utility of IRT for scale development and refinement.

Another area in which IRT analyses greatly facilitated the development of the SSIS SEL*b* was in assessing items and scales for potential bias and fairness. Our DIF analyses singled out SSIS SEL items that would be most likely to lead to unfair or inequitable decisions made about children, which allowed us to exclude these items from the SSIS SEL*b*. Indeed, the vast majority of retained SSIS SEL*b* items did not have evidence of DIF. Yet, due to content considerations and limited options among scales with more items evidencing DIF, we retained several items with small DIF for the SSIS SEL*b*. Fortunately, because the magnitude of DIF was small in all cases, the influence of these retentions is likely to be very minor (Fig. 1) and constrained to relatively small ranges of construct continua. Specifically, since interpretation would likely utilize raw scores, DIF could lead to differences of up to 1 point between White and Non-white students for the SSIS SEL*b*-PP Self-Management and Social Awareness scales. These differences would potentially favor White children at average-high ranges of these constructs and favor Non-white children at lower levels of these constructs. Such differences are very minor, yet users should be aware of them and exercise appropriate caution. Use of the SSIS SEL*b* Composite scores, which aggregate scores across the five scales, for decision making would further limit the negative impact of this DIF and thus is recommended for use in applied practice.

It is also interesting to note that more instances of DIF arose in evaluating the SSIS SEL-P than the SSIS SEL-T. This finding could have arisen due to the fact that the sample used to conduct these analyses was larger for the SSIS SEL-P than the SSIS SEL-T. Alternatively, these differences may reflect cultural differences between subgroups as parents were presumably more likely to be of the same racial/ethnic group as their children. In contrast, teachers were likely more racial homogenous, in line with national estimates (e.g., Hussar et al., 2020) although we did not have reported teacher demographic data to confirm this. It is also interesting that DIF did not appear to be related across corresponding items on the SSIS SEL-T and SSIS SEL-P. That is, if an item was flagged for DIF on the SSIS SEL-T it did not necessarily evidence DIF on the SSIS SEL-P. Clearly, there is much more to be learned regarding a nuanced understanding of social behavior and expectations for social behavior across racial/ethnic groups and this represents an important area for future research.

Next, validity analyses generally supported the score validity of the SSIS SEL*b*. Scale intercorrelations and correlations with the SSRS were largely in line with expectations and similar to equivalent statistics calculated with the SSIS SEL. These findings provide initial validity evidence for scores from the SSIS SEL*b*, although future studies are needed to further validate the measure. One notable finding we did not anticipate concerned the SSRS Internalizing scale, which had lower magnitude correlations with SSIS SEL*b* scales than expected. This finding could possibly be due to the difficulty of assessing internalizing symptoms in preschool students, especially by third party raters (Poulou, 2015). The Internalizing scale of the SSRS has the lowest reliability coefficients of any SSRS scale, which could attenuate validity coefficients in the current study. Despite this attenuation, it is notable that the validity coefficients for the SSIS SEL*b* Relationship Skills scale and the SSRS Internalizing scale were higher in magnitude than for any other SSIS SEL*b* scale correlation with this SSRS scale across both the SSIS SEL*b*-TP and the SSIS SEL*b*-PP. This finding does provide some support for the validity of scores from the SSIS SEL*b* as behaviors reflected in the Relationship Skills scale (e.g., *Interacts well with other children; Starts conversations with peers*) would likely be differentially impacted negatively by internalizing symptoms such as withdrawal relative to other SEL skills (e.g., *Is well-behaved when unsupervised*).

Generally, reliability and validity evidence was similar when calculated using the SSIS SEL and SSIS SEL*b*. The largest differences between validity coefficients were for the Relationship Skills SSIS SEL and SSIS SEL*b* scales in which 64% of validity coefficients were statistically significantly different (generally weaker) when calculated with the SSIS SEL*b*. This is likely due in part to the fact the SSIS SEL Relationship Skills scale has more items ($N = 13$ and 14 for the SSIS SEL–Teacher and Parent respectively) than most SSIS SEL scales, so more content was cut to reduce the SSIS SEL*b* to four items. The SSIS SEL Relationship Skills scale does not have the most items of any SSIS SEL scale, however (the SSIS SEL Self-Management scale has similarly high numbers of items), which implies that the Relationship Skills scale is possibly more heterogeneous in its content than other SSIS SEL scales. Thus, potential users should note that although the SSIS SEL*b* likely assesses the core of the relationship skills construct, supplementation of this domain with the full SSIS SEL or another measure might be beneficial. Another source of evidence that differed when calculated with the SSIS SEL relative to the SSIS SEL*b* regarded intercorrelations, which were notably stronger when calculated with the SSIS SEL compared with the SSIS SEL*b*. In some ways, this evidence could be taken to support the discriminant validity of the SSIS SEL*b* scales, as others (Panayiotou, Humphrey, & Wigelsworth, 2019) have criticized the SSIS SEL and

similar measures for intercorrelations that might be considered excessively strong. Indeed, this problem appears with other SEL assessments such as the Devereux Student Strengths Assessment (DESSA; LeBuffe, Shapiro, & Naglieri, 2009). For example, Doromal, Cottone, & Kim (2019) conducted a CFA of the DESSA and found interscale correlations ranging from .81 to .90 (Median = .86). Thus, it is possible that CASEL SEL domains are exceedingly difficult for raters to conceptually distinguish. Future theoretical and empirical work is needed to explore this possibility.

4.1. Limitations and future research

Despite the general success of the current study, there are several important limitations. First, there were several methodological limitations that should be noted. Specifically, the sample size for our IRT analyses involving the SSIS SEL–Teacher was small relative to many applications of IRT. To bolster the small sample size of the standardization sample, we incorporated cases collected during the SSIS standardization that were not ultimately included in the standardization sample, which rendered the full dataset less representative than the original standardization data. Further, data were originally gathered in 2006–2007 as part of the SSIS standardization, and may not be representative of the current preschool population. Future research replicating our analyses with larger, current, and more diverse samples is warranted.

Another notable limitation of the current project regarded the lack of appropriate content to generate a SSIS SEL*b* Self-Awareness scale. This limitation is one that is not unique to the SSIS SEL for preschool children (Humphrey et al., 2011), but one that is important given the presence of state standards for Self-Awareness in all 50 states and Washington D.C. (Eklund et al., 2018). It is possible that the self-awareness construct is too internalized for external raters to reasonably assess because preschool children are not developmentally ready to exhibit behavioral indicators of self-awareness amenable to rating scale technology (McKown, 2017). Yet, a successful augmentation of the SSIS SEL*b* would greatly increase its utility and applicability and should be attempted. There were also some limitations of our DIF analyses. Specifically, due to sample size requirements, we had to collapse Nonwhite children (Black, Hispanic, and Other) into one group. This allowed the use of DIF to further refine the SSIS SEL*b*, but it might obscure important differences between these subgroups. Future work with larger samples is warranted to evaluate DIF at a more fine-grained level.

Finally, the current analyses represent development work and *initial* validation, and future research is needed to continue the ongoing process of validation (AERA, APA, & NCME, 2014). Specifically, future research should validate the SSIS SEL*b* when administered as a standalone measure with independent samples (Smith, McCarthy, & Anderson, 2000). Beyond additional data it will be important to examine different sources of validity evidence. Perhaps most importantly, the SSRS and SSIS SEL*b* measure similar constructs and thus correlations between these measures represent strong convergent validity data. Current evidence does not, however, address the discriminant validity of SSIS SEL*b* scores. This is particularly important because it is possible that in the development process only the most general SSIS SEL items were selected, omitting more specific indicators of SEL skills that differentiate scores from dissimilar constructs. Because of this possibility, future research addressing this limitation represents a key next step in the ongoing validation process of the SSIS SEL*b*.

Furthermore, future research is needed in several domains to continue to strengthen the evidence for specific applications of the SSIS SEL*b*. First, to support identification of children in need of further assessment/intervention, diagnostic accuracy analysis or ROC curve studies would be very useful. Considering prior research indicating that SEL skills promote school readiness (Denham et al., 2014), indicators of early academic skill proficiency, social relationships, and/or emotional regulation at kindergarten entry might be especially relevant for prediction and cut-score generation. Next, to support the use of the SSIS SEL*b* for periodic monitoring of children's growth in SEL skills, future analyses evaluating the change sensitivity of SSIS SEL*b*-TP and SSIS SEL*b*-PP scores is warranted. These directions are only two possibilities in the ongoing validation process of the SSIS SEL*b*, but they represent important further directions to build on the strong foundation of evidence for interpretation and use of the SSIS SEL*b*-Preschool Forms.

4.2. Implications for research and practice

The SSIS SEL*b* holds promise for both research and practice. First, with regard to research, there currently are few preschool measures that are content aligned with the CASEL framework. Considering the ubiquity of this model in state standards and preschool educational practice, it is clear that much more research is needed regarding preschool service delivery from within the CASEL framework. Future research should link CASEL-aligned assessment with interventions specifically designed for preschool settings and should examine the impacts of improving children's SEL skills in this developmental period. The SSIS SEL*b* should promote such intervention validation work, although researchers should exercise caution until more evidence establishes the change-sensitivity of scores from these measures. The current investigation also highlights several important advantages of using IRT for scale development/refinement. Indeed, IRT affords several advantages that are particularly beneficial for preschool level measurement work including better understanding the construct levels at which instruments provide precise measurement (which would be very helpful for identifying and ameliorating floor/ceiling effects commonly found in preschool samples) and advantages for longitudinally scaling assessments for more valid and precise measurement during this dynamic developmental stage (e.g., McDermott, Rikoon, & Fantuzzo, 2014). Given these benefits, researchers should consider IRT for future measurement development for preschoolers.

There are several potential applications for the SSIS SEL*b*. The measure would potentially function well for universal assessment purposes, as well as periodic progress monitoring and group-based assessment, although specific evidence is needed for these applications before they can be fully supported. There are two particularly important features of the SSIS SEL*b* that increase its utility. First, the SSIS SEL*b* is a multirater instrument designed to maximize content alignment across informants. This is critical considering the fact that prior to preschool, parents have the most information to provide of any rater regarding their children's SEL skills. Thus, gathering data efficiently from multiple raters can promote more integrated and effective service provision. Such assessment may be best timed as children enter and exit preschool to plan SEL focused service delivery as children transition between schools. Next, the SSIS SEL*b* is brief and efficient, which is especially important for teachers. In the context of many modern assessment applications such as universal screening and progress monitoring, teacher raters are required to complete assessments for many children (i.e., in universal screening) or many times on the same child (i.e., in progress monitoring). Thus, establishing brief measures that do not sacrifice technical quality is especially important considering currently prominent assessment applications. The SSIS SEL*b* meets this need and should promote CASEL-aligned preschool assessment. Finally, a key strength of the SSIS system is that it is explicitly linked with evidence-based prevention and intervention programs (e.g., DiPerna et al., 2018). Yet, these programs have not yet been extended to preschool populations. Thus, if these programs were adapted for preschool populations, the SSIS SEL*b*-Preschool Forms

could likely be integrated into such a system. Regardless of whether such development work occurs, the SSIS SEL*b*-Preschool Forms could likely be integrated with current preschool SEL-focused prevention programs to comprehensively and efficiently promote SEL skills of all preschoolers.

5. Conclusions

In sum, the SSIS SEL*b* holds promise to promote SEL-focused research and practice for preschool children. Currently, there are very few measures that, like the SSIS SEL*b*, align with the broadly prominent CASEL framework, connect with the widely used SSRS and SSIS, inform evidence-based intervention, and are highly efficient. Despite these advantages, there are important areas for further development, most notably determining if it is possible to create a developmentally-appropriate Self-Awareness scale for the SSIS SEL*b*, extending the research base to support specific applied uses of the SSIS SEL*b*, and replicating the current findings with recent, representative samples, and alternate validity measures. Such future work will be important for the SSIS SEL*b* and similar measures to more fully support comprehensive and high quality SEL service delivery for all preschool children.

Authors' contribution

Christopher J. Anthony: conceptualization, methodology, formal analysis, data curation, writing – original draft, writing – review & editing, visualization, supervision. Stephen N. Elliott: conceptualization, methodology, writing – original draft, writing – review and editing. James C. DiPerna: conceptualization, methodology, writing – review and editing. Pui-Wa Lei: conceptualization, methodology, writing – review and editing.

Conflict of interest

The SSIS SEL Brief Scales–Preschool Forms are published by SAIL-CoLab, and all authors receive royalties from the distribution of these measures.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Anderson, N., Schlueter, J. E., Carlson, J. F., & Geisinger, J. F. (Eds.). (2016). *Tests in print IX*. Lincoln, NE: Buros Center for Testing. ISBN 978-0-910674-65-2

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. http://dx.doi.org/10.1016/j.jsp.2015.12.005

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. http://dx.doi.org/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. http://dx.doi.org/10.1007/s12310-018-9254-7

Anthony, C. J., Elliott, S. N., DiPerna, J. C., & Lei, P.-W. (2020). The SSIS SEL Brief Scales?Student Form: Initial development and validation. *School Psychology*, *35*(4), 277–283. http://dx.doi.org/10.1037/spq0000390

Bierman, K. L., Greenberg, M. T., & Abenavoli, R. (2016). *Promoting social and emotional learning in preschool: Programs and practices that work*. Edna Bennet Pierce Prevention Research Center, Pennsylvania State University.

Bierman, K. L., & Motamedi, M. (2015). SEL programs for preschool children. In J. A. Durlak, C. E. Domitrovich, R. P. Weissberg, & T. P. Gullotta (Eds.), *Handbook of social and emotional learning: Research and practice* (pp. 135–150). New York: Guilford Press.

CASEL. (2015). *Effective social and emotional learning programs: Middle and high school*. Chicago, IL: Author.

Cai, L., Thissen, D., & du Toit, S. H. C. (2019). *IRTPRO for Windows [Computer software]*. Lincolnwood, IL: Scientific Software International.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Denham, S. A., Bassett, H. H., Zinsser, K., & Wyatt, T. M. (2014). How preschoolers' social emotional learning predicts their early school success: Developing theory-promoting, competency-based assessments. *Infant and Child Development*, *23*(4), 426–454. http://dx.doi.org/10.1002/icd.1840

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), 1. http://dx.doi.org/10.1037/edu0000191

Doromal, J. B., Cottone, E. A., & Kim, H. (2019). Preliminary validation of the teacher-rate DESSA in a low-income, kindergarten sample. *Journal of Psychoeducational Assessment*, *37*(1), 40–54. http://dx.doi.org/10.1177/0734282917731460

Dusenbury, L., Dermody, C., & Weissberg, R. P. (2018). *State scorecard scan*. Retrieved from https://casel.org/state-scan-scorecard-project-2/.

Eklund, K., Kilpatrick, K. D., Kilgus, S. P., & Haider, A. (2018). Children, research, and public policy. *School Psychology Review*, *47*(3), 316–326. http://dx.doi.org/10.17105/SPR-2017.0116.V47-3

Elias, M. J., & Moceri, D. C. (2012). Developing social and emotional aspects of learning: The American experience. *Research Papers in Education*, *27*(4), 423–434. http://dx.doi.org/10.1080/02671522.2012.690243

Elliott, S. N., & Gresham, F. M. (2017). *SSIS SEL edition screening/progress monitoring scales*. Bloomington, MN: Pearson Assessments.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. London, UK: Lawrence Erlbaum Associates.

Frey, J. R., Elliott, S. N., & Gresham, F. M. (2011). Preschoolers' social skills: Advances in assessment for intervention using social behavior ratings. *School Mental Health*, *3*(4), 179–190. http://dx.doi.org/10.1007/s12310-011-9060-y

Gresham, F. M., Cook, C. R., Collins, T., Dart, E., Rasetshwane, K., Truelson, E., & Grant, S. (2010). Developing a change-sensitive brief behavior rating scale as a progress monitoring tool for social behavior: An example using the Social Skills Rating System—Teacher Form. *School Psychology Review*, *39*(3), 364–379.

Gresham, F. M., Elliott, S. N., & Black, F. L. (1987). Factor structure replication and bias investigation of the Teacher Ratings of Social Skills. *Journal of School Psychology*, *25*, 81–92. http://dx.doi.org/10.1016/0022-4405(87)90063-X

Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system*. Circle Pines MN: AGS.

Gresham, F. M., & Elliott, S. N. (2017). *SSIS SEL edition rating forms*. Bloomington, MN: Pearson Assessments. http://dx.doi.org/10.1007/978-1-4614-6435-8_102184-1

Gresham, F. M., Elliott, S. N., Byrd, S., Wilson, E., & Cassidy, K. (2018). Cross-informant agreement of children's social emotional skills: An investigation of ratings by teachers, parents, and students from a nationally representative sample. *Psychology in the Schools*, *55*, 208–223. http://dx.doi.org/10.1002/pits.22101

Gresham, F. M., Elliott, S. N., Metallo, S., Byrd, S., Erickson, M., Cassidy, K., et al. (2018). Psychometric fundamentals of the Social Skills Improvement System Social Learning Edition Rating Forms. *Assessment for Effective Intervention*, http://dx.doi.org/10.1177/1534508418808598

Halle, T. G., & Darling-Churchill, K. E. (2016). Review of measures of social and emotional development. *Journal of Applied Developmental Psychology*, *45*, 8–18. http://dx.doi.org/10.1016/j.appdev.2016.02.003

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*(4), 617–637. http://dx.doi.org/10.1177/0013164410382896

Hussar, B., Zhang, J., Hein, S., Wang, K., Roberts, A., Cui, J., et al. (2020). *The Condition of Education 2020 (NCES 2020-144)*. Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2020144.

Jones, S., Bailey, R., Brush, K., & Nelson, B. (2018). Introduction to the Taxonomy Project: Tools for Selecting & Aligning SEL Frameworks. Establishing Practical Social-Emotional Competence Assessments Work Group. Retrieved from: https://measuringsel.casel.org/wp-content/uploads/2019/02/Frameworks-C.1.pdf.

LeBuffe, P. A., & Shapiro, V. B. (2004). Lending "strength" to the assessment of preschool social-emotional health. *The California School Psychologist*, *9*(1), 51–61. http://dx.doi.org/10.1007/BF03340907

LeBuffe, P. A., Shapiro, V. B., & Naglieri, J. A. (2009). *The Devereux Student Strengths Assessment (DESSA)*. Lewisville, North Carolina: Kaplan.

Lee, I. A., & Preacher, K. J. (2013). *Calculation for the test of the difference between two dependent correlations with one variable in common*. (Computer software).

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130. http://dx.doi.org/10.1037/1082-989X.1.2.130

McDermott, P. A., Rikoon, S. H., & Fantuzzo, J. W. (2014). Tracing children's approaches to learning through Head Start, kindergarten, and first grade: Different pathways to different outcomes. *Journal of Educational Psychology*, *106*(1), 200–213. http://dx.doi.org/10.1037/a0033547

McKown, C. (2017). *Social-emotional assessment, performance, and standards*. pp. 157–178. The Future of Children.

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. http://dx.doi.org/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. http://dx.doi.org/10.1037/spq0000334

Muthén, L.K. and Muthén, B.O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.
National Reading Panel (US), NICHHD (US), National Reading Excellence Initiative, National Institute for Literacy (US), & United States Department of Health. (2000). *Report of the National Reading Panel*. NICHHD, NIH.
Panayiotou, M., Humphrey, N., & Wigelsworth, M. (2019). An empirical basis for linking social and emotional learning to academic performance. *Contemporary Educational Psychology*, *56*, 193–204. http://dx.doi.org/10.1016/j.cedpsych.2019.01.009
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*, *18*(1), 25–34. http://dx.doi.org/10.1016/j.jval.2014.10.005
Poulou, M. S. (2015). Emotional and behavioural difficulties in preschool. *Journal of Child and Family Studies*, *24*(2), 225–236. http://dx.doi.org/10.1007/s10826-013-9828-9
Reeve, B. B., Hays, R. D., Chang, C. H., & Perfetto, E. M. (2007). Applying item response theory to enhance health outcomes assessment. *Quality of Life Research*, *16*, 1–3. http://dx.doi.org/10.1007/s11136-007-9220-6
Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48. http://dx.doi.org/10.1146/annurev.clinpsy.032408.153553
Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. http://dx.doi.org/10.1037/a0029315
Salvia, J., Ysseldyke, J. E., & Witmer, S. (2016). *Assessment in Special and Inclusive Education (13th ed.)*. Boston, MA: Cengage Learning.
Salvia, J., Ysseldyke, J. E., & Witmer, S. (2017). *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 Monograph Supplement*.
Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, *12*(1), 102–111. http://dx.doi.org/10.1037/1040-3590.12.1.102
Snyder, T. D., de Brey, C., & Dillow, S. A. (2019). *Digest of education statistics 2017, NCES 2018-070*. National Center for Education Statistics.
Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251.
Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*(1), 3–46. http://dx.doi.org/10.1177/1094428114553062
Tourangeau, K., Nord, C., Lê, T., Pollack, J. M., & Atkins-Burnett, S. (2006). *Early childhood longitudinal study, kindergarten class of 1998-99 (ECLS-K), combined user's manual for the ECLS-K fifth-grade data files and electronic codebooks (NCES 2006-032)*. Washington, DC: US Department of Education, National Center for Education Statistics.
Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Vaden-Kiernan, N., Blaker, L., & Najarian, M. (2019). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) User's Manual for the ECLS-K:2011 Kindergarten–Fifth Grade Data File and Electronic Codebook, Public Version (NCES 2019-051)*. U.S. Department of Education. Washington, DC: National Center for Education Statistics.
Wang, H. T., Sandall, S. R., Davis, C. A., & Thomas, C. J. (2011). Social skills assessment in young children with autism: A comparison evaluation of the SSRS and PKBS. *Journal of Autism and Developmental Disorders*, *41*(11), 1487–1495. http://dx.doi.org/10.1007/s10803-010-1175-8
Wollersheim Shervey, 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*, *32*(3), 414–421. http://dx.doi.org/10.1037/spq0000203