Journal of School Psychology 55 (2016) 57–69

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

Maximizing measurement efficiency of behavior rating scales using Item Response Theory: An example with the Social Skills Improvement System — Teacher Rating Scale☆

Christopher J. Anthony *, James C. DiPerna, Pui-Wa Lei

The Pennsylvania State University, USA

ARTICLE INFO

Article history:
Received 7 October 2014
Received in revised form 9 December 2015
Accepted 17 December 2015
Available online 18 January 2016

Keywords:
Social skills
Problem Behaviors
Measurement efficiency
Item Response Theory
Polytomous models

ABSTRACT

Measurement efficiency is an important consideration when developing behavior rating scales for use in research and practice. Although most published scales have been developed within a Classical Test Theory (CTT) framework, Item Response Theory (IRT) offers several advantages for developing scales that maximize measurement efficiency. The current study provides an example of using IRT to maximize rating scale efficiency with the Social Skills Improvement System — Teacher Rating Scale (SSIS — TRS), a measure of student social skills frequently used in practice and research. Based on IRT analyses, 27 items from the Social Skills subscales and 14 items from the Problem Behavior subscales of the SSIS — TRS were identified as maximally efficient. In addition to maintaining similar content coverage to the published version, these sets of maximally efficient items demonstrated similar psychometric properties to the published SSIS — TRS.

1. Introduction

Difficulties in the area of social skills have been linked with numerous problems including juvenile delinquency (Roff, Sell, & Golden, 1972), social isolation (Chung et al., 2007; Matson & Boisjoli, 2007), and school drop-out (Parker & Asher, 1987). In addition, social skills demonstrate moderate relationships with academic achievement (Wentzel, 1993; Malecki & Elliott, 2002). As a result of their relationship with student success, social skills have been identified as an important educational outcome. Currently, all 50 states have explicitly specified standards related to social–emotional functioning, and 96% of states have standards for social–emotional preschool development (DiPerna, Bailey, & Anthony, 2014). Further, learning in schools is laden with social and emotional dimensions due to the necessity that students function within the social environment of a school (Zins, Weissberg, Wang, & Walberg, 2004). As such, efficient and accurate assessment of social skills is an important focus for education professionals and researchers.

Several rating scales have been developed to measure the social skills of children from preschool through high school. Examples include the School Social Behavior Scale-2 (Merrell, 2002) and the Walker–McConnell Scales of Social Competence and School Adjustment (Walker & McConnell, 1995). The Social Skills Rating System (SSRS; Gresham & Elliott, 1990) has been widely used to

☆ The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A090438 to The Pennsylvania State University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.
* Corresponding author at: Department of Educational Psychology, Counseling, and Special Education, The Pennsylvania State University, 125 CEDAR Building, University Park, PA 16802, USA. Tel: +1 406 210 3527. fax: +1 814 865 7066.
E-mail address: cja5171@psu.edu (C.J. Anthony).
Action Editor: Michelle Demaray

http://dx.doi.org/10.1016/j.jsp.2015.12.005
0022-4405/

assess children's social skills in both research and practice. Gresham, Elliott, Vance, and Cook (2011) reported that the SSRS was used in 127 studies published in 50 different peer-reviewed journals and 53 doctoral dissertations between 2003 and 2008. The revision of the SSRS, the Social Skills Improvement System — Teacher Rating Scales (SSIS — TRS) features a broader conceptualization of several domains of social skills (Gresham et al., 2011) and scores with psychometric properties similar to those of the original version.

Given the widespread use of behavior rating scales to assess social skills in research and practice, the efficiency with which a rating scale measures social skills is of critical importance. Teachers often are asked to complete multiple assessments as part of a comprehensive evaluation process, which further stretches the limited amount of non-instructional time they have available during the day. Similarly, the time required to complete rating scales is challenging in a research context as well. School-based research participants (often teachers) may have to complete several rating forms per class at several time points throughout the year. Although they may be compensated for these activities, inefficiency of measurement and unnecessary time burdens can add to teacher stress and lead to incomplete data or withdrawal from study participation. As a result, many researchers shorten existing measures to minimize time required of participants (e.g., Duckworth, Quinn, & Tsukayama, 2012). Given the importance of the reliability and validity of these scores for substantive research, development and evaluation of shorter or streamlined versions of social skills measures is crucial.

1.1. Test theory and measurement efficiency

Most behavior rating scales, including the SSIS — TRS, have been developed within a Classical Test Theory (CTT) framework. Briefly, CTT is based on the assumption that the observed score produced by a measure is equal to the sum of two parts: the true score, which reflects a student's ability, and measurement error, which results from any systematic or random factor not related to the construct of interest (Suen, 2008). One commonly examined source of such error is that associated with differences among items. This type of error is quantified and measured indirectly by examining the internal consistency of a measure, which refers to how well items on a measure are related (Barchard, 2010). A major focus of scale development in a CTT context is the maximization of internal consistency (Streiner, 2003). Although internal consistency is undeniably important, overemphasizing it can be problematic for several reasons (Briggs & Cheek, 1986). First, this emphasis encourages inclusion of a larger number of items, which increases Cronbach's α provided items are positively related to the total score. Second, maximizing internal consistency can result in the inclusion of highly similar items within a measure. Indeed, Streiner (2003) noted that Cronbach's α levels that exceed .90, a threshold often used for evaluating technical adequacy for individual decisions (e.g., Salvia, Ysseldyke, & Bolt, 2010) more likely indicate unnecessary redundancy rather than optimal internal consistency. As such, rating scales developed in a CTT framework may display some inefficiency, the elimination of which would result in more parsimonious and rater-friendly measures.

Efficiency of measurement is well addressed by another theoretical framework under which rating scales can be developed, Item Response Theory (IRT; Lord, 1980; Hambleton & Swaminathan, 1985). Broadly, IRT refers to a series of latent trait methods used to assess item functioning and estimate latent trait score. A major advantage of IRT in the context of measurement efficiency is its facilitation of graphical evaluation of item performance (Edelen & Reeve, 2007). As part of IRT procedures, information curves are produced displaying the level of information (akin to precision of measurement) produced by a particular item across varying levels of an underlying ability or trait (e.g., social skills; Hambleton & Swaminathan, 1985). This advantage allows rating scale developers to identify items that provide desired amounts of information at desired levels of an underlying ability or trait. For example, if a researcher is interested in developing a measure to be used primarily to identify children who have low social skills (i.e., a measure in which high precision at low levels of social skills is desired), the researcher could select items that give high information at lower levels of social skills. As a result, fewer items would be required to achieve a desired level of precision at the targeted social skill levels (because items that function well at higher social skill levels would be excluded). Such item evaluation also can be conducted on item parameter estimates produced by IRT, such as an item's discrimination (which is positively related to information) and location on the latent trait scale at which an item is discriminating (often referred to as location or difficulty). As a whole, these procedures allow researchers to identify and eliminate items that do not function as desired relative to the measure's primary purpose, thus maximizing efficiency of measurement.

In addition to the general utility of IRT in improving measurement efficiency, assessing model assumptions can be helpful in identifying sources of measurement inefficiency. Alongside the assumption of unidimensionality, two major assumptions tested within the IRT framework are the assumptions of local independence (Hambleton & Swaminathan, 1985) and functional form (Toland, 2014). Local independence requires items to be mutually independent (or uncorrelated in a weaker form of the assumption) after the latent trait(s) that they purport to measure are controlled. A violation of this assumption indicates that some items share common variance that is not due to the intended latent traits. Among various potential causes of local dependence, item redundancy is of particular interest for streamlining measures. For example, similar or slightly different wordings of essentially the identical items can result in local dependence. Violations can also be due to incorrectly specifying the dimensionality of the scale or other incorrect model specifications, however, so it is important to examine item content for redundancy and examine other potential causes of local dependence.

Furthermore, the assumption of functional form states that the empirical data or distribution roughly follows the function specified by the IRT model chosen to analyze the data (De Ayala, 2009). This assumption is examined through reviewing Option Characteristic Curves (OCCs) as well as calculating item- and model-level fit statistics. Examining OCCs can be especially helpful in the context of increasing measurement efficiency (Toland, 2014). Often, poorly functioning items or response categories can be

identified by visual analysis of OCCs. For example, option categories that are unused or not ordered properly can be modified or eliminated through this analysis. Thus, IRT methodology offers researchers several advantages for maximizing the measurement efficiency of rating scales.

1.2. Rationale and purpose

Given the many time constraints placed on teachers in schools today, researchers and practitioners need measures of social skills that are both efficient and psychometrically sound. The SSIS — TRS is widely used in practice and research; however it was developed within a CTT framework and may benefit from IRT analyses to maximize its efficiency. As such, the primary purpose of this study was to provide an example of how to use IRT to maximize the efficiency of rating scale measures by identifying a Set of Maximally Efficient Items (SMI) for the Social Skills and Problem Behaviors scales of the SSIS — TRS. A secondary purpose was to examine internal consistency, stability, and criterion and convergent validity relationships for the scores from the SMI relative to those reported for the published version of the SSIS — TRS. These analyses were conducted to provide insight regarding the comparability of the SMIs and the SSIS — TRS and to examine the level of reliability of scores from the SMIs and the validity of specific interpretations of SMI scores.

2. Method

2.1. Participants

Participants included 299 first-graders and 484 second-graders with an overall mean age of 7.27 years ($SD = 0.93$). There were slightly (1%) more girls than boys and only 2% spoke a language other than English, of which 0.5% spoke Spanish. With regard to race, 67% of the sample was White, 15% were Black/African American, 6% were Hispanic or Latino, 2% identified as other, 1.3% were Asian, 0.6% were Multiracial, and 0.5% were Native Hawaiian/Pacific Islander (0.5% of students were identified as "Unknown" on this variable).[1] In addition, 7% of the sample had an identified disability that qualified them for special education services, over half of which had qualified as having a Learning Disability. Participants were in 55 different classrooms across six schools in two school districts (one rural, one small urban). Demographic data were not available for 6 teachers, but of the 49 remaining teachers, 88% were female. With regard to race, 98% of the teachers identified as White and 2% identified as African American. Teachers' primary language was English, and they had taught for an average of 15.42 years ($SD = 9.27$) overall and 10.61 years ($SD = 9.14$) at their current grade.

2.2. Measures

2.2.1. Social Skills Improvement System — Teacher Rating Scale (SSIS — TRS)

The Social Skills and Problem Behavior scales of the SSIS — TRS (Gresham & Elliott, 2008a) were the focus of this study. The Social Skills scale includes the subscales of Communication (7 items), Cooperation (6 items), Assertion (7 items), Responsibility (6 items), Empathy (6 items), Engagement (7 items), and Self-Control (7 items); and the Problem Behaviors domain includes the subscales of Externalizing (12 items), Bullying (5 items), Hyperactivity/Inattention (7 items), and Internalizing (7 items). These domains are rated on a 4-point Likert-type scale from 1 (*never*) to 4 (*almost always*).

Evidence for reliability of the scores from the SSIS — TRS is reported in the test manual (Gresham & Elliott, 2008b). For the 5- to 12-year-old age group, Cronbach's α was .97 for the Social Skills scale and ranged from .83 to .92 (median = .90) for the Social Skills subscales. For the Problem Behaviors scale, Cronbach's α was .95 and ranged from .78 to .93 (median = .88) for the Problem Behaviors subscales. Similarly, stability coefficients (2–87 day intervals) for the Social Skills scale and subscales ranged from .68 to .85 (median = .81). For the Problem Behaviors scale and subscales, stability coefficients ranged from .76 to .86 (median = .84). With regard to convergent validity, scales and subscales of the SSIS — TRS correlated as expected with several measures (Gresham & Elliott, 2008b) of related constructs such as the Externalizing Problems subscale of the Behavioral Assessment System for Children-Second Edition (BASC-2; Reynolds & Kamphaus, 2004) and the Socialization scale of the Vineland Adaptive Behavioral Behavior Scales (Vineland-II; Sparrow, Cicchetti, & Balla, 2005). Additionally, scores from the SSIS — TRS accurately differentiated clinical and non-clinical samples for a number of groups including children with Autism, Attention Deficit/Hyperactivity Disorder, and Intellectual Disability. Finally, other studies (e.g., Gresham et al., 2011; Gresham, Elliott, Cook, Vance, & Kettler, 2010) using different samples have similarly provided evidence for the convergent validity, internal consistency, and interrater reliability of scores from the SSIS — TRS.

2.2.2. Academic Competence Evaluation Scales (ACES)

The Motivation and Engagement subscales from the ACES Teacher Record Form (DiPerna & Elliott, 2000) were administered as part of this study. These subscales were included because they have been shown to demonstrate positive relationships with social skills (e.g., DiPerna, Volpe, & Elliott, 2002, 2005). Items on the ACES are rated on a 5-point Likert-type scale ranging from 1 (*never*) to 5 (*almost always*). Internal consistency coefficients of the Motivation and Engagement subscales were .98 and .94 respectively

[1] Roughly 7% of cases were missing data on race. As such, percentages do not sum to 100% for this variable.

for Kindergarten through second grade students in the standardization sample. The same coefficients for the current sample were .98 and. 96 respectively. Furthermore, stability coefficients (2–3 week intervals) were .96 and .92 for the Motivation and Engagement subscales in a subsample of the standardization sample (DiPerna & Elliott, 2000). Scores from the Motivation and Engagement subscales of the ACES have also displayed large to medium[2] correlations with both reading (DiPerna et al., 2002) and mathematics (DiPerna et al., 2005) achievement. Furthermore, scores from these ACES subscales have demonstrated medium to large relationships with measures of student social behavior (DiPerna & Elliott, 2000). Finally, the theoretical factor structure of the ACES was supported by factor analysis by the authors (DiPerna & Elliott, 2000). Overall, there is evidence for reliability and validity of scores from the ACES.

2.2.3. STAR Reading and Mathematics

Given the documented positive relationships between social functioning and academic achievement (e.g., Wentzel, 1993; Malecki & Elliott, 2002) STAR Reading and STAR Mathematics (Rennaissance Learning, Inc., 2007) also were used to examine the evidence of criterion related validity of SMI scores in this study. The STAR assessments are computer-adaptive assessments that measure reading and mathematics skills for students in Grades 1–12. As evidence of reliability, internal consistency coefficients range from .89 to .91 (median = .90) and stability coefficients from .82 to .89 (median = .83) for students in the first to fifth grade (U.S. Department of Education, 2010) for STAR Reading scores. Similarly, STAR Mathematics scores have demonstrated internal consistency coefficients ranging from .79 to .83 (median = .81) and stability coefficients ranging from .73 to .79 (median = .74) for students in the first through fifth grades (U.S. Department of Education, 2010). With regard to validity, meta-anlayses have shown that scores from both the STAR Reading and the STAR Mathematics assessments highly correlate with scores from other standardized achievement tests. Specifically, validity coefficients ranging from .71 to .73 (median = .72) for the STAR Reading and from .63 to .65 (median = .64) for the STAR Mathematics (U.S. Department of Education, 2010).

2.3. Procedure

Data were drawn from an efficacy trial of the Social Skills Improvement System: Classwide Intervention Program (SSIS — CIP; Elliott & Gresham, 2007). In this larger study, classrooms were randomly assigned to treatment (SSIS — CIP implementation) and control (business as usual) conditions. Prior to conducting the study, parent and teacher consent, as well as student assent were obtained. Teachers then completed the SSIS — TRS and ACES subscales[3] in the fall (baseline) and spring (post-SSIS — CIP implementation) for a subsample of students in their classroom. During each time point, all teachers completed the SSIS — TRS followed by the ACES, and research staff monitored teachers' progress to ensure both measures were completed in their entirety for all participating students in their classroom. In addition, participating students were administered the Star Reading and Mathematics assessments during the same data collection period when the teachers completed the SSIS — TRS and ACES. Trained research assistants facilitated administration of the STAR via laptops. To ensure the design of the efficacy trial did not impact the results of the current study, only fall (baseline) scores were used for the IRT and concurrent validity analyses. In addition, the spring scores from the subsample of participants in the no-treatment comparison condition were the only scores included in the stability analyses.

2.4. Polytomous IRT models

Ordered polytomous IRT models developed for items such as those found on behavior ratings scales were used in this study. These IRT models assume that response categories are ordered such that endorsing a higher category reflects a higher trait level than endorsing a lower category. They also estimate the probability of endorsing a particular response category as a function of the level of an underlying latent trait (often referred to as theta or θ), a category location parameter (called threshold or intercept parameter), and where applicable, an item discrimination parameter. When plotted as a curve, the function is called an Option Characteristic Curve (OCC). Each item has as many OCCs as categories (e.g., an item utilizing a 5-point Likert-type scale would have 5 OCCs). The threshold parameters are related to the crossing points between adjacent OCCs (Ostini & Nering, 2006). Participants with trait levels above the threshold have higher probability of endorsing the adjacent higher category than the adjacent lower one. The discrimination parameter (commonly called the *a* parameter) indicates the extent to which the item discriminates between individuals with different trait levels and it is functionally related to the amount of item information. Item information can be summed across items that are included in a scale to form the scale or test level information. Test information is inversely related to standard error of measurement (i.e., the more information, the lower the standard error of measurement), which can be different for individuals with different trait levels. The additive nature of item information is one of the reasons why IRT can be used conveniently to create test forms to meet specific needs (e.g., a certain amount of measurement precision within certain range of trait levels).

Polytomous IRT models are different with respect to constraints imposed on item parameters. The Partial Credit Model (PCM; Masters, 1982) assumes that items on a scale are equally discriminating. As such, discrimination parameters are constrained to be

[2] All descriptive labels of correlation coefficients correspond to Cohen's (1988) classification labels.

[3] Because the brief SSIS — TRS Academic Competence Scale assesses similar constructs to the more comprehensive ACES Motivation and Engagement subscales, only the SSIS — TRS Social Skills and Problem Behaviors subscales were administered. Similarly, the ACES Academic Skills scale was not administered due to the inclusion of the STAR measures.

1 for all items (an alternate version of PCM only constrains them to be equal but not necessarily equal to 1). Threshold parameters are allowed to be different from each other and across items. An extension of the PCM, the Generalized Partial Credit Model (GPCM; Muraki, 1992) removes the equality constraint on the discrimination parameters and estimates a separate discrimination parameter for each item. The Graded Response Model (GRM; Samejima, 1969) is similar to GPCM in that the same numbers of item discrimination and threshold parameters are estimated. These two models (the GRM and GPCM) differ in the way probabilities for response categories are defined and calculated. As a result of their similarity, GRM and GPCM tend to produce similar estimates and model fit statistics (Ostini & Nering, 2006).

3. Results

3.1. Tests of assumptions

3.1.1. Assumption of unidimensionality

Several steps were undertaken to determine whether IRT analyses were appropriate. First, although there are multidimensional IRT models (Hambleton & Swaminathan, 1985), most IRT applications assume unidimensionality of the scale analyzed (as do most applications of CTT; Amtmann et al., 2010). Because IRT analyses were conducted at the subscale level, confirmatory factor analyses were employed to assess the assumption of unidimensionality for each subscale. MPlus (Muthén & Muthén, 2012) was used for these analyses. Support for unidimensionality was assessed by overall fit index cutoffs recommended by Hu and Bentler (1999). Sets of items often will not meet strict criteria of unidimensionality. In these cases, if the items are found to be "essentially" unidimensional, IRT analyses are considered appropriate. Essential unidimensionality indicates that although a set of items may not meet strict criteria for unidimensionality, the presence of other dimensions does not significantly affect the parameter estimates generated by IRT analyses under the assumption of unidimensionality (Amtmann et al., 2010). Thus, for scales with some evidence of poor fit, follow up Exploratory Factor Analyses (EFA) were conducted and evaluated according to criteria established by Reeve et al. (2007) to examine essential unidimensionality. Specifically, EFA analyses were considered to support essential unidimensionality if the first factor accounted for 20% or more of the overall variance or if the ratio of the eigenvalues of the first factor to the second factor exceeded 4.

Although most fit indices for one-factor CFA model indicated adequate fit according to a priori criteria, a few relevant fit indices did not meet these standards. The lack of statistically significant modification indices for any scale, however, indicated that inter-item correlations were explained by the one-factor model. Furthermore, all item loadings were sufficiently high (>.6) in all models. Factor loadings ranged from .62 to .97 across all subscales. On follow-up EFAs, all subscales showed strong evidence of essential unidimensionality with a ratio of the eigenvalues of the first factor to the second factor ranging from 7 to 71.14. Further, the first factor accounted for more than 20% of the variance for all subscales. In sum, according to procedures set forth by Reeve et al. (2007), these analyses supported essential unidimensionality, and unidimensional IRT models were determined to be appropriate for all subscales.

3.1.2. Assumption of local independence

Another assumption of IRT is local independence, which requires pair-wise item responses to be uncorrelated conditional on the latent trait. There are multiple ways of assessing the assumption of local independence; however, for the purposes of this project, local dependence χ^2 values (test statistics for the null hypothesis that conditional covariance between item pairs = 0) produced by IRTPRO (Cai, Thissen, & du Toit, 2011) were evaluated. Values > 10 were considered to indicate excessive local

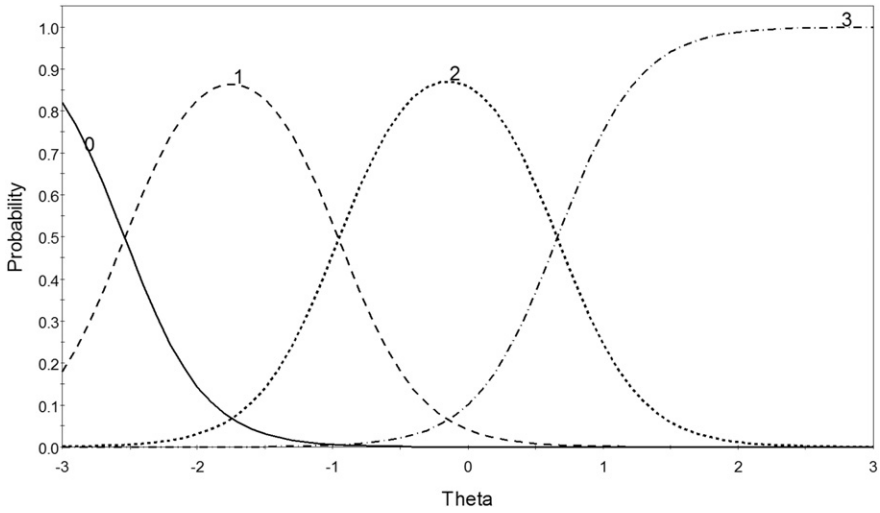

Fig. 1. Sample Option Characteristic Curve for the first item of the Assertion Subscale Set of Maximally Efficient Items (SMI).

dependence (Cai et al., 2011). Likely due to the fact that the SSIS — TRS subscales were not initially developed within an IRT framework, there was evidence of local dependence in several of the subscales when analyzed with the GRM. Items that contributed to local dependence were typically similar items (e.g., asking whether a child is inattentive or is easily distracted). These shortcomings were addressed during item analysis and selection of maximally efficient items.

3.1.3. Assumption of functional form

A final foundational assumption of IRT analyses is functional form, which is that the data follow the function specified by the model chosen to analyze the data (De Ayala, 2009). For example, in the context of the GRM, the assumption of functional form would imply that all threshold parameters are ordered and there is a common slope for each item (Toland, 2014). This assumption is often tested through model and item fit statistics (discussed in the following section) and visual inspection of an item's OCCs to ensure that each item's categorical rating system is functioning as intended. Specifically, in a well-functioning categorical rating system, the lowest category is most likely to be endorsed at low levels of theta, the highest category is most likely to be endorsed at high levels of theta, and intermediary categories are more likely to be endorsed than the preceding category as one moves from lower to higher values along the theta continuum (see Fig. 1 for an example). All SSIS — TRF OCCs were examined and found to function as expected. This is further reflected by the sequential ordering of all threshold parameters for each subscale (Table 1). Along with the item level and overall model fit statistics discussed below, these OCCs and threshold parameter estimates support the assumed functional form.

3.2. Model fit

3.2.1. Item level fit

Because IRT is an item level modeling approach, fit at the item level is usually established prior to examining overall model fit. There are various methods of examining item level fit such as the G^2 statistic (McKinley & Mills, 1985) and the Q1 statistic (Yen, 1981). For the current study, a generalization of Orlando and Thissen's (2000, 2003) S-χ^2 statistic was emphasized due to its advantages for polytomous models (Kang & Chen, 2008), large samples, and small numbers of items (Orlando & Thissen, 2003). This statistic assesses the level of similarity between expected (i.e. predicted by the model chosen for the analysis) and observed response frequencies by category. This statistic follows a χ^2 distribution and a statistically significant result indicates a difference between predicted and observed response frequencies. Given the potential for Type I error inflation due to the many subscales analyzed and the many statistical comparisons conducted, the a priori alpha level was set to .01 for each subscale. Furthermore, Bonferroni adjustments were applied within subscale for these analyses (e.g., .002 for the Responsibility subscale which has 6 items, i.e., .01 divided by 6). Also, because χ^2 test would be inaccurate when expected cell frequencies are sparse (e.g., <5 for many cells), the S−χ^2 statistics were calculated twice with minimum expected cell frequency of 1 (Orlando & Thissen, 2000) and 5 (Kang & Chen, 2008). Items were considered to fit poorly if they were flagged across these two minimum expected frequency cell values. Results reported were obtained from the GRM, the model eventually chosen for all analyses based on model fit statistics.

Based on these procedures, 14 of the 77 analyzed items displayed poor fit. These items were distributed across five subscales: Communication (7 items), Assertion (3 items), Empathy (2 items), Self-Control (1 items), and Cooperation (1 item). For these items, the full observed versus expected frequency tables produced by IRTPRO were examined, and long strings of over- or under-estimation were taken to indicate misfit of the item model. A random pattern of over- and under-estimation was taken to indicate adequate item level fit. Evaluation of these tables indicated that, despite statistically significant χ^2 statistics for several items, retention of most of them would not be problematic. Nevertheless, item fit was considered during item selection, resulting in the inclusion of only 4 items with statistically significant χ^2 values in the final SMIs (3 on the Communications subscale [for

Table 1
Descriptive statistics for item discrimination (a) and threshold (b) parameters by subscale of the SSIS — TRS.

Subscale	*a*				b_1				b_2				b_3			
	M	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Social skills																
Communication	2.03	0.71	1.18	3.02	−2.41	0.20	−2.80	−2.23	−1.21	0.16	−1.39	−0.91	0.20	0.19	−0.03	0.56
Cooperation	2.41	0.69	1.66	3.37	−2.05	0.26	−2.41	−1.67	−0.82	0.27	−1.15	−0.38	0.48	0.28	0.22	1.01
Assertion	1.44	0.61	0.84	2.29	−2.44	0.54	−2.96	−1.30	−0.69	0.33	−0.93	−0.08	0.79	0.32	0.39	1.27
Responsibility	2.62	1.10	1.47	4.24	−2.29	0.24	−2.58	−2.02	−1.10	0.18	−1.33	−0.92	0.25	0.20	−0.01	0.43
Empathy	2.99	1.16	1.62	4.58	−2.46	0.32	−2.97	−2.10	−1.12	0.27	−1.46	−0.68	0.37	0.21	0.22	0.76
Engagement	2.17	0.44	1.41	2.64	−2.53	0.25	−2.76	−2.03	−1.12	0.29	−1.53	−0.61	0.29	0.26	−0.05	0.74
Self Control	2.24	0.55	1.78	3.37	−2.12	0.22	−2.53	−1.89	−0.98	0.23	−1.34	−0.76	0.50	0.28	0.09	0.79
Problem behaviors																
Externalizing	1.96	0.40	1.17	2.65	0.52	0.51	−0.52	1.24	1.59	0.38	0.91	2.18	2.53	0.30	2.13	3.10
Bullying	2.86	0.99	1.91	4.48	0.94	0.22	0.62	1.13	1.91	0.14	1.67	2.00	2.78	0.15	2.62	2.98
Hyperactivity/Inattention	1.82	0.52	1.01	2.55	0.14	0.71	−0.48	1.25	1.23	0.72	0.50	2.30	2.42	0.91	1.48	3.84
Internalizing	1.70	0.57	1.00	2.76	0.45	0.40	−0.34	0.98	1.73	0.37	1.44	2.46	2.86	0.31	2.58	3.44

Note. SSIS — TRS = Social Skills Improvement System — Teacher Rating Scale; all parameters are presented on the Normal Ogive Scale.

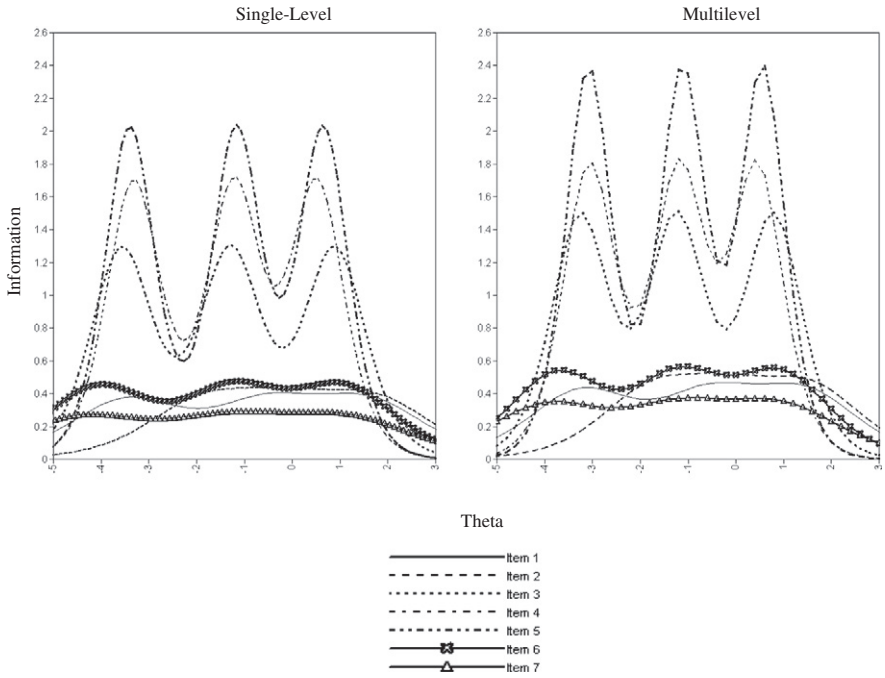

Fig. 2. Example graphs of item information curves for the Assertion Subscale for single-level and multilevel models. Note. Both models were computed with MPlus so that figure layout is the same to facilitate comparison. The single-level model results are similar between IRTPRO and MPlus.

which no item had nonsignificant χ^2 values] and 1 item on the Self-Control subscale). This approach was taken so as not to overemphasize item fit relative to other important considerations such as content coverage, overall model fit, and the presence of local dependence in final SMIs.

3.2.2. Overall model fit

Next, overall model fit was assessed and compared for various IRT models. For the purposes of comparison, models are either considered nested or non-nested. A model is considered to be nested within a more complex model when it constrains one or more parameters of the more complex model. For example, the PCM is considered to be nested within the GPCM because the PCM constrains the discrimination parameter of the GPCM. For this project, all models were compared with the −2 log likelihood value, the Akaike Information Criterion (Akaike, 1974) and Bayesian Information Criterion (Schwarz, 1978) statistics. For these statistics, lower values indicate better fit. Nested models were additionally compared with a −2 log likelihood difference test which references a χ^2 distribution with degrees of freedom equivalent to the difference in number of estimated parameters of the two compared models. The following polytomous models were fitted with IRTPRO: (a) a PCM, (b) a PCM with a common *a* parameter, (c) a GPCM, and (d) a GRM. For each subscale, the GRM was found to provide the best model fit. Descriptive statistics for the GRM parameter estimates by subscale can be found in Table 1.

Next, analyses were conducted to determine the necessity of accounting for the multilevel structure of the data. This step of the analysis was especially important because students were not only nested within classroom (mean cluster size = 14.27), but also within rater, which could result in important differences between single-level and multilevel models due to the rater effect. A typical first step of evaluating the need for multilevel modeling includes an examination of sample Intraclass Correlation Coefficients (ICCs; Stapleton, 2013). These ICCs were calculated according to procedures described by Raykov and Marcoulides (2015), which accounts for categorical data in ICC calculation, and ranged from .08 to .49 (median = .27) across items. In light of the high ICCs, both single and multilevel models[4] were conducted and compared. All single level models were conducted with IRTPRO (Cai et al., 2011), and multilevel models allowing class-level latent trait estimates to vary were conducted using MPlus (Muthén & Muthén, 2012). Because the object of measurement was student level Social Skills and Problem Behaviors, results of single-level models were compared with Level 1 results from multilevel models. Although there were minor differences across model type, substantive results (e.g., which items produced superior information curves, location of high information)

[4] A full multilevel model was conducted rather than design-based modeling (Stapleton, 2013) because the former best accounted for the structure of the data. Readers interested in further information regarding the application of multilevel IRT models should consult Fox (2004, 2013).

Table 2
Number of retained and omitted Items, correlations between SSIS — TRS and SMI scales and subscales, correlations between SMI scales and subscales and corresponding omitted items, and RMSEA for SSIS — TRS and SMI subscales.

	# of items		Correlation between		RMSEA	
Scale/subscale	Retained	Omitted	SSIS — TRS & SMI	SMI & omitted items	SSIS — TRS	SMI
Social skills	27	19	.95	.96		
Communication	3	4	.90	.81	.08	.06
Cooperation	4	2	.95	.83	.05	.05
Assertion	4	3	.91	.71	.07	.10
Responsibility	4	2	.93	.80	.08	.07
Empathy	4	2	.93	.83	.07	.09
Engagement	4	3	.93	.82	.06	.08
Self Control	4	3	.91	.86	.11	.07
Problem behaviors	14[a]	11	.94	.93		
Externalizing	7	5	.94	.86	.05	.05
Bullying	4	1	.93	.73	.04	.06
Hyperactivity/Inattention	5	2	.95	.82	.07	.05
Internalizing	4	3	.92	.79	.05	.06

Note. SSIS — TRS = Social Skills Improvement System — Teacher Rating Scale; SMI = Set of Maximally Efficient Items; RMSEA = Root Mean Square Error of Approximation.

[a] The sum is greater than the total number of items retained and omitted for the Problem Behaviors scale because several Problem Behaviors items are included on multiple scales in the SSIS — TRS.

were equivalent. An example of comparison curves can be found in Fig. 2. Because substantive results were equivalent, single level models are reported.[5]

Each model also produces several overall model fit statistics. Model fit statistics are produced by comparing the pattern of actual responses of an item or scale to the pattern of responses that would be predicted by each model. The primary overall model fit statistic used for this project was the Root Mean Square Error of Approximation (RMSEA). RMSEA values were taken to indicate acceptable fit if they did not exceed .10 (MacCallum, Browne, & Sugawara, 1996). After the selection of the SMIs, the GRM was fit again with each SMI subscale. For the SMIs, RMSEA ranged from .04 to .11 (median = .07) for SSIS — TRS subscales and from .05 to .10 (median = .06) for SMI subscales. The SMI item selection process resulted generally in no changes or decreases in RMSEA, although one scale (Internalizing) increased in RMSEA by .01, three scales (Empathy, Engagement, and Bullying) increased by .02, and one scale (Assertion) increased by .03 (Table 2). Thus, overall model fit for each of these SMI subscales was acceptable according to a priori criteria.

3.2.3. Item analysis

There are several possible criteria that can be used to identify maximally efficient items to inform short form development. In line with the emphasis on efficiency, the first step of this project was to eliminate items that contributed to local dependency due to content or wording similarity. To do so, when two items were found to be locally dependent they were compared and the poorer performing item was eliminated. The main performance criterion was the item information curves across values of theta. Items were favored if they maximized information at lower levels of theta because behavior rating scales such as the SSIS — TRS are often used for evaluating the social behaviors of students experiencing difficulty. (Items were favored if they maximized information at higher levels of theta for Problem Behavior subscales). Another criterion was to ensure minimum loss of information across the theta scale and to preserve the shape of the original test information function as much as possible. A final consideration was item level fit and ensuring good overall fit in each SMI subscale (based on the RMSEA value of the scale). Because the models used in this study assume unidimensionality, IRT analyses were conducted on each subscale of the SSIS — TRS rather than the entire scale. This also ensured that all subscales would be preserved.

As a result of item and scale analyses according to procedures outlined previously, 27 Social Skills items and 14 Problem Behaviors items were identified as maximally efficient (while retaining sufficient content coverage). All SMI subscales had large correlations of .90 or greater with their SSIS — TRS counterparts (Table 2) indicating that the forms were highly similar. Also, to examine the construct overlap between the retained and omitted item sets, scores from the SMI scales and subscales were correlated with scores from the corresponding sets of omitted SSIS — TRS items (Table 2). Score correlations between the retained and omitted sets of items for the Social Skills and Problem Behavior scales were .96 and .93 respectively. Correlations ranged from .71 to .86 (median = .83) for the Social Skills subscales and from .73 to .86 (median = .81) for the Problem Behavior subscales.

3.2.4. Evidence for reliability/precision of SMI scores

Information curves were compared to examine the measurement precision of scores from the SMIs compared to scores from the SSIS — TRS. For each scale and subscale, the shape of the information curve for the SMI of each scale and subscale was similar

[5] Although no differences were observed between single- and multilevel models in this study, researchers must be aware of potential data dependency when using the SSIS — TRS and properly take it into account if observed.

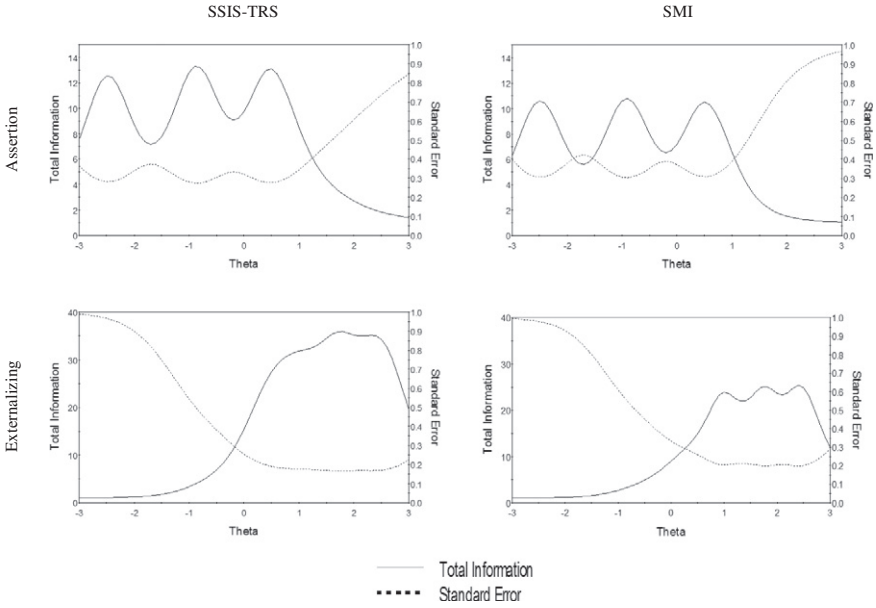

Fig. 3. Graphs of the total information curves and standard error curves across levels of theta for the Assertion and Externalizing Subscales of the Social Skills Improvement System — Teacher Rating Scale and the Set of Maximally Efficient Items (SMI).

to the shape of the information curve for the full set of items on the SSIS — TRS. Test information curves for both the SSIS — TRS and the SMI of two representative subscales are shown in Fig. 3. Not surprisingly, information levels were slightly lower (as evidenced on the Y axes of the graphs in Fig. 3) for several of the SMI scales and subscales; however, overall loss of information was minimal considering almost half the items were eliminated through the item selection process.

Cronbach's α was computed to examine internal consistency of the SSIS — TRS and SMI scales and subscales. As shown in Table 3, all were above the .80 minimum criterion commonly used to evaluate screening measures (Salvia et al., 2010) and were minimally different than SSIS — TRS α levels. The difference between SSIS — TRS and SMI α levels was 0.01 for the Social Skills scale and ranged from 0.01 to 0.03 (median = 0.02) for the Social Skills subscales. The same difference was 0.03 for the Problem Behaviors scale and ranged from 0.02 to 0.06 (median = 0.04) for the Problem Behaviors subscales. In addition, test-retest stability coefficients were calculated between SSIS — TRS scores[6] collected in the fall and spring of the same year for both SMI and the SSIS — TRS (Table 3). The average time interval between data collection was 4.5 months (*SD* = .29 months). As noted previously, stability coefficients were only calculated for children in the control group (n = 383) of the larger study. These stability coefficients were .69 for both the SMI Social Skills and Problem Behaviors scales and ranged from .60 to .72 (median = .65) for the SMI Social Skills subscales and from .57 to .71 (median = .66) for the SMI Problem Behaviors subscales. The difference between SSIS — TRS and SMI stability coefficients was 0.10 for the Social Skills scale and ranged from 0.07 to 0.13 (median = .10) for Social Skills subscales. The same difference was 0.11 for the Problem Behaviors scale and ranged from .09 to .13 (median = .10) for Problem Behaviors subscales.

3.3. Evidence for relationship of SMI scores to other variables

The final analyses examined the criterion related and convergent validity relationships of scores from the SMI and related constructs. First, the SMI Social Skills scale and subscales demonstrated small positive relationships with the STAR measures, and the SMI Problem Behaviors scale and subscales demonstrated small negative relationships with these measures. Correlations of scores from the SMIs with STAR Reading were .26 for the SMI Social Skills scale and ranged from .11 to .28 (median = .23) for the subscales. The SMI Problem Behaviors scale and STAR Reading correlation was −0.24, and correlations across Problem Behavior subscales ranged from −.23 to −.14 (median = −.21). A similar pattern emerged with concurrent relationships for STAR Mathematics scores. The SMI correlations with STAR mathematics were .36 for the SMI Social Skills scale and ranged from .19 to .36 (median = .30) for the subscales. For SMI Problem Behaviors scale, the correlation was −.28 and ranged from −.28 to −.15 (median = −.24) for SMI Problem Behaviors subscales (Table 4).

In addition, the SMIs and the SSIS — TRS demonstrated expected convergent relationships with the ACES Motivation and Engagement subscales (Table 5). For the ACES Motivation scores, correlations were .70 and −.49 with the SMI Social Skills and SMI Problem Behaviors scales respectively. Further, ACES Motivation correlations ranged from .50 to .74 (median = .59) with

[6] Mean scores of SSIS — TRS and SMI scales were used for calculation of stability coefficients and validity coefficients.

Table 3
Internal consistency (Cronbach's α) and stability coefficients for the SSIS — TRS and the SMI scales and subscales.

Scale/subscale	Cronbach's α		Stability coefficients	
	SSIS—TRS	SMI	SSIS — TRS [95% CI[a]]	SMI [95% CI]
Social skills	.98	.97	.79 [.75, .83]	.69 [.63, .74]
Communication	.91	.88	.74 [.69, .78]	.65 [.59, .70]
Cooperation	.93	.91	.79 [.75, .83]	.72 [.67, .77]
Assertion	.86	.85	.71 [.66, .76]	.61 [.54, .67]
Responsibility	.92	.90	.78 [.74, .82]	.66 [.60, .71]
Empathy	.94	.92	.71 [.66, .76]	.64 [.58, .70]
Engagement	.93	.91	.74 [.69, .78]	.66 [.60, .71]
Self Control	.93	.90	.77 [.73, .81]	.64 [.58, .70]
Problem Behaviors	.94	.91	.80 [.76, .83]	.69 [.63, .74]
Externalizing	.93	.89	.80 [.76, .83]	.70 [.64, .75]
Bullying	.91	.89	.70 [.64, .75]	.57 [.50, .63]
Hyperactivity/Inattention	.90	.84	.80 [.76, .83]	.71 [.66, .76]
Internalizing	.88	.86	.72 [.67, .77]	.62 [.55, .68]

Note. SSIS — TRS = Social Skills Improvement System — Teacher Rating Scale; SMI = Set of Maximally Efficient Items.
[a] All confidence intervals reported were calculated by conducting Fisher's Z transformation on each original coefficient, adding and subtracting 1.96 * the standard error ($1 / \sqrt{[N-3]}$) to the Z transformed value, and transforming the resulting upper and lower bounds back to their original scale.

SMI Social Skills subscales and from −.53 to −.26 (median = −.40) with SMI Problem Behaviors subscales. Again, a similar pattern was seen with the ACES Engagement scores as the criterion. These correlations were .68 for the SMI Social Skills scale and −.42 for the SMI Problem Behaviors scale. ACES Engagement correlations ranged from .47 to .67 (median = .59) with the SMI Social Skills subscales and from −.46 to −.19 (median = −.33) with the SMI Problem Behaviors subscales.

Finally, as further evidence that SMI scores function similarly to SSIS — TRS scores, differences between these correlations when calculated with SSIS — TRS scores and SMI scores were minimal. For STAR Reading and Mathematics scores, these differences ranged from −.02 to .04 (median = 0) for Social Skills scales and subscales and from −.03 to 0 (median = −.03) for Problem Behaviors scales and subscales. For ACES scores, differences ranged from −.02 to .08 (median = .05) for Social Skills scales and subscales and from −.07 to 0 (median = −.04) for Problem Behaviors scales and subscales. Overall, differences between SSIS — TRS and SMI correlations with related constructs were within measurement error as evidenced by overlapping confidence intervals for each corresponding SMI and SSIS — TRS correlation (Tables 4 and 5).

4. Discussion

4.1. Summary of major findings

The SSIS — TRS is arguably the most prominent measure of school children's social skills in research and practice. As demonstrated in the current study, the SMI of the SSIS — TRS identified by IRT item analyses produced scores that exhibited similar psychometric properties to those of the full-length SSIS — TRS. With regard to evidence for the reliability of SMI scores, TIFs

Table 4
Correlations between SSIS — TRS and SMI scales and subscales and STAR Reading and Math Scores.

Scale/subscale	STAR Reading		STAR Math	
	SSIS — TRS [95% CI[a]]	SMI [95% CI]	SSIS — TRS [95% CI]	SMI [95% CI]
Social skills	.26 [.19, .33]	.26 [.19, .33]	.36 [.29, .42]	.36 [.29, .42]
Communication	.25 [.18, .32]	.23 [.16, .30]	.33 [.26, .39]	.32 [.25, .38]
Cooperation	.26 [.19, .33]	.28 [.21, .35]	.35 [.28, .41]	.36 [.29, .42]
Assertion	.18 [.11, .25]	.20 [.13, .27]	.28 [.21, .35]	.30 [.23, .37]
Responsibility	.27 [.20, .34]	.28 [.21, .35]	.36 [.29, .42]	.36 [.29, .42]
Empathy	.13 [.06, .20]	.11 [.04, .18]	.22 [.15, .29]	.19 [.12, .26]
Engagement	.23 [.16, .30]	.21 [.14, .28]	.34 [.27, .40]	.30 [.23, .37]
Self Control	.21 [.14, .28]	.23 [.16, .30]	.27 [.20, .34]	.27 [.20, .34]
Problem Behaviors	−.26 [−.33, −.19]	−.24 [−.31, −.17]	−.31 [−.37, −.24]	−.28 [−.35, −.21]
Externalizing	−.22 [−.29, −.15]	−.19 [−.26, −.12]	−.26 [−.33, −.19]	−.23 [−.30, −.16]
Bullying	−.15 [−.22, −.08]	−.14 [−.21, −.07]	−.15 [−.22, −.08]	−.15 [−.22, −.08]
Hyperactivity/Inattention	−.25 [−.32, −.18]	−.22 [−.29, −.15]	−.31 [−.37, −.24]	−.28 [−.35, −.21]
Internalizing	−.23 [−.30, −.16]	−.23 [−.30, −.16]	−.26 [−.33, −.19]	−.25 [−.32, −.18]

Note. SSIS — TRS = Social Skills Improvement System — Teacher Rating Scale; SMI = Set of Maximally Efficient Items.
[a] All confidence intervals reported were calculated by conducting Fisher's Z transformation on each original coefficient, adding and subtracting 1.96 * the standard error ($1 / \sqrt{[N-3]}$) to the Z transformed value, and transforming the resulting upper and lower bounds back to their original scale.

Table 5
Correlations between SSIS — TRS and SMI scales and subscales and ACES Motivation and Engagement scores.

Scale/subscale	ACES Motivation		ACES Engagement	
	SSIS — TRS [95% CI[a]]	SMI [95% CI]	SSIS — TRS [95% CI]	SMI [95% CI]
Social skills	.74 [.71, .77]	.70 [.66, .73]	.71 [.67, .74]	.68 [.64, .72]
Communication	.67 [.63, .71]	.62 [.57, .66]	.68 [.64, .72]	.63 [.59, .67]
Cooperation	.74 [.71, .77]	.74 [.71, .77]	.57 [.52, .62]	.59 [.54, .63]
Assertion	.58 [.53, .62]	.52 [.47, .57]	.73 [.70, .76]	.67 [.63, .71]
Responsibility	.70 [.66, .73]	.68 [.64, .72]	.57 [.52, .62]	.58 [.53, .62]
Empathy	.57 [.52, .62]	.50 [.45, .55]	.55 [.50, .60]	.48 [.42, .53]
Engagement	.65 [.61, .69]	.59 [.54, .63]	.72 [.68, .75]	.64 [.60, .68]
Self Control	.54 [.49, .59]	.54 [.49, .59]	.46 [.40, .51]	.47 [.41, .52]
Problem Behaviors	−.55 [−.60, −.50]	−.49 [−.54, −.43]	−.45 [−.50, −.39]	−.42 [−.48, −.40]
Externalizing	−.48 [−.53, −.42]	−.43 [−.49, −.37]	−.31 [−.37, −.24]	−.29 [−.35, −.20]
Bullying	−.27 [−.33, −.20]	−.26 [−.32, −.19]	−.19 [−.26, −.12]	−.19 [−.26, −.10]
Hyperactivity/Inattention	−.60 [−.64, −.55]	−.53 [−.58, −.48]	−.41 [−.47, −.35]	−.37 [−.43, −.30]
Internalizing	−.43 [−.49, −.37]	−.37 [−.43, −.31]	−.50 [−.55, −.45]	−.46 [−.51, −.40]

Note. SSIS — TRS = Social Skills Improvement System — Teacher Rating Scale; SMI = Set of Maximally Efficient Items; ACES = Academic Competence Evaluation Scales.

[a] All confidence intervals reported were calculated by conducting Fisher's Z transformation on each original coefficient, adding and subtracting 1.96 * the standard error ($1 / \sqrt{[N-3]}$) to the Z transformed value, and transforming the resulting upper and lower bounds back to their original scale.

were largely similar between SSIS — TRS and SMI scales and subscales. Furthermore, internal consistency estimates exceeded the Cronbach's α standard of .80, a common criterion for individual planning and intervention (Salvia et al., 2010).

In contrast, most of the stability coefficients for SMI were more attenuated by the item analysis procedures. There are several important considerations for the interpretation of this finding. First, the time interval between data collection was quite long. Specifically, the longest interval reported in the SSIS manual (Gresham & Elliott, 2008a) was 87 days, which is almost 2 months shorter than the average interval reported in these analyses. Furthermore, the constructs measured are amenable to change (indeed, the SSIS — TRS instructions indicate for raters to base their rating on the past 2 months of student behavior). For these reasons, although most of the observed stability coefficients are lower than their SSIS — TRS counterparts, the observed stability coefficients should be interpreted in light of these considerations.

Although internal consistency and stability coefficients are reported, it is important to note the limitations of these indices compared to information curves produced in IRT analyses. As mentioned above, one of the major advantages of IRT is that test developers are able to examine measurement precision across levels of theta. As demonstrated in Fig. 3, measurement precision often is not uniform across levels of theta, and, as a result, reliability coefficients conceal variation in score reliability. Thus, although it is appropriate to consider traditional internal consistency and stability coefficients, IRT analyses provide additional information about score precision, as evidenced by these analyses.

With regard to validity evidence, scores from the SSIS — TRS and SMIs of the Social Skills scale and subscales demonstrated small to medium positive correlations with STAR Reading and Mathematics scores. Further, scores from the SSIS — TRS and SMIs of the Problem Behaviors scale and subscales had small negative correlations with STAR Reading and Mathematics scores. These coefficients are smaller in magnitude than others reported in the literature (Malecki & Elliott, 2002). Although smaller in magnitude, these patterns were similar to those observed in other research on the concurrent relationship between social skills and academic achievement. The SSIS—TRS and SMI Social Skills scale and subscales also demonstrated mostly large (only two validity coefficients fell below $r = .5$) positive correlations with the ACES Motivation and Engagement scores. Further, scores from the Problem Behaviors scale and subscales showed a small to medium negative correlation with the ACES Motivation and Engagement scores. These patterns are similar to previous research on the relationship between SSRS and the ACES scores (DiPerna & Elliott, 2000).

As further evidence of SMI score validity, all concurrent relationship coefficients were extremely similar (within measurement error as evidenced by overlapping confidence intervals of corresponding correlation coefficients in Tables 4 and 5) when calculated with scores from the SMI and the SSIS — TRS. The SMIs accounted for almost the same amount of variance in criterion variables as accounted for by their SSIS — TRS counterparts. Specifically, SMI scales and subscales accounted for an average of 0.3% less variance in STAR scores than their SSIS — TRS counterparts and 3.7% less variance in ACES scores. Thus, although 29 items were eliminated through the process of IRT analyses, the relationships between scores of the SMI and academic achievement, motivation, and engagement were virtually identical to the relationships between the SSIS — TRS scores and academic achievement, motivation, and engagement.

4.2. Limitations and research directions

There are several limitations to the current study. First, all calculated reliability and concurrent relationship coefficients are preliminary, as the items may function differently when completed in the context of a standalone short form (as opposed to being completed among the larger pool of SSIS — TRS items). Furthermore, although scores from the SMI were highly correlated with scores from the SSIS — TRS, these correlations reflect scores based on many of the same items rated by the same raters at the

same time. Also, four items included in the SMIs did not display adequate fit to the data, although three of these came from a subscale with no adequately fitting items. With regard to overall fit, several of the SMI subscales (Assertion, Empathy, and Engagement) demonstrated minimally adequate fit to the data as evidenced by RMSEA values between .08 and .10 (MacCallum et al., 1996). In addition, 5 scales increased in RMSEA as a result of item selection, although these increases were minimal. Also, as item selection procedures favored items with higher information at lower levels of theta for Social Skills subscales and higher levels of theta for Problem Behaviors subscales, the resulting SMIs would likely not be efficient if used with students with high Social Skills or low Problem Behaviors ratings.

In addition, although the correlations between the SMI, STAR measures, and ACES subscales provide some evidence of validity, no measure of children's Social Skills and Problem Behaviors was included in the study. Such examination should be undertaken in future studies. Also, the magnitude of validity coefficients between both the SMIs and SSIS — TRSs and measures of achievement were slightly lower than other estimates of the relationship between social skills and academic achievement (e.g., Malecki & Elliott, 2002). This was especially true for reading validity coefficients. Despite these differences, the direction and magnitude of all SSIS — TRS validity coefficients still supported the convergent validity of both forms of the SSIS — TRS analyzed. Nevertheless, these results indicate the need for further validity evidence of the scores from both the SMIs developed in this project and the SSIS — TRS.

As a result of these limitations, there are several important directions for future research. First, if a short form of the SSIS — TRS was created based on the SMI from this study, it would need to be field tested to see if it functions similarly to the reduced set of items identified in the current study. Future studies also should assess the structure of such an IRT informed short form with Confirmatory Factor Analysis to examine internal structure of the scores from the SMI. Also, to address evidence based on test content, expert review of the item content is necessary to ensure sufficient coverage of the social skill constructs intended to be measured. Finally, follow-up studies should examine the acceptability and utility of any IRT informed short form of the SSIS — TRS.

4.3. Conclusion

The findings from this study have potentially significant implications for research and practice. The SSIS — TRS manual states that the published version of the SSIS — TRS can be completed in 15 to 20 min (Gresham & Elliott, 2008b). In contrast, a version based on the SMI would be able to be completed in approximately half that time (7–10 min). Currently, many researchers must use longer instruments that, though their scores may demonstrate technically adequate psychometric characteristics, require more time from research participants. As evidenced by the similar concurrent coefficients produced by the SSIS — TRS and SMI identified in this study, the SMI yields scores that provide a good estimate of the domains measured by the SSIS — TRS in about half the time needed to complete the full version of the measure. Such time savings can decrease the burden on research participants. Similarly, a short form could be beneficial to practitioners in the context of multi-tiered service delivery systems in which measures of varying depth are useful at different tiers of service.

In addition to these potential implications, the process described in this study provides an example of the use of IRT to maximize the measurement efficiency of rating scale measures. Given the time constraints under which many educational professionals work and the limited resources available to educational researchers, measurement efficiency should be a primary consideration for test developers. The IRT procedures described in this paper are especially relevant for this goal and should be considered for the development of new measures or the modification and improvement of current measures.

Overall, the SMI developed in this study has potential to inform a short form measure of social skills for both educational practitioners and researchers. Given further development and validation studies, such an IRT informed short form could serve as an efficient, technically adequate tool for measuring students' Social Skills and Problem Behaviors in the classroom setting and for educational research. Further, the methodology used for this project demonstrates the value of IRT in the development of behavior rating scale measures in general, and short forms of existing behavior rating scale measures in particular. As evidenced by this study, such methodology can help researchers maximize the measurement efficiency of behavior rating scales for optimal use in educational research and practice.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. http://dx.doi.org/10.1109/TAC.1974.1100705.

Amtmann, D., Cook, K. F., Jensen, M. P., Chen, W., Choi, S., Revicki, D., ... Lai, J. (2010). Development of a PROMIS item bank to measure pain interference. *Pain*, *150*, 173–182. http://dx.doi.org/10.1016/j.pain.

Barchard, K. (2010). In N. J. Salkind (Ed.), *Internal consistency reliability. Encyclopedia of Research Design.*. (pp. 616–620). Thousand Oaks, CA: SAGE Publications, Inc. http://dx.doi.org/10.4135/9781412961288.n191.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the evaluation of personality scales. *Journal of Personality*, *54*, 106–148. http://dx.doi.org/10.1111/j.1467-6494.1986.t00391.x.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows [Computer software].* Lincolnwood, IL: Scientific Software International.

Chung, K. M., Reavis, S., Mosconi, M., Drewry, J., Matthews, T., & Tasse, M. J. (2007). Peer-mediated social skills training program for young children with high-functioning autism. *Research in Developmental Disabilities*, *28*, 423–436. http://dx.doi.org/10.1016/j.ridd.2005.05.002.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: Guilford.

DiPerna, J. C., & Elliott, S. N. (2000). *Academic Competence Evaluation Scales.* San Antonio, TX: The Psychological Corporation.

DiPerna, J. C., Bailey, C. G., & Anthony, C. J. (2014). Broadband screening of academic and social behavior. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Identification, implementation, and interpretation.* Washington DC: APA Books.

DiPerna, J. C., Volpe, R. J., & Elliott, S. N. (2002). A model of academic enablers and elementary reading/language arts achievement. *School Psychology Review*, *31*, 298–312.
DiPerna, J. C., Volpe, R. J., & Elliott, S. N. (2005). A model of academic enablers and mathematics achievement in the elementary grades. *Journal of School Psychology*, *43*, 379–392. http://dx.doi.org/10.1016/j.jsp.2005.09.002.
Duckworth, A. L., Quinn, P. D., & Tsukayama, E. (2012). What No Child Left Behind leaves behind: The roles of IQ and self-control in predicting standardized achievement test scores and report card grades. *Journal of Educational Psychology*, *104*, 439–451. http://dx.doi.org/10.1037/a0026280.
Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*, 5–18. http://dx.doi.org/10.1007/s11136-007-9198-0.
Elliott, S. N., & Gresham, F. M. (2007b). *Social skills improvement system: Classwide intervention program guide.* Bloomington, MN: Pearson Assessments.
Fox, J. -P. (2004). Applications of multilevel IRT modeling. *School Effectiveness and School Improvement*, *15*, 261–280.
Fox, J. -P. (2013). Introduction to multilevel IRT modeling. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory. Vol. 1*, . Chapman and Hall/CRC Press (Chapter 24).
Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system.* Minneapolis, MN: NCS Pearson.
Gresham, F. M., & Elliott, S. N. (2008a). *Social skills improvement system: Teacher rating scales.* Bloomington, MN: Pearson Assessments.
Gresham, F. M., & Elliott, S. N. (2008b). *Social skills improvement system: Rating scales manual.* Minneapolis, MN: Pearson Assessments.
Gresham, F. M., Elliott, S. N., Cook, C. R., Vance, M. J., & Kettler, R. (2010). Cross-informant agreement for ratings for social skill and problem behavior ratings: An investigation of the social skills improvement system — Rating scales. *Psychological Assessment*, *22*, 157–166. http://dx.doi.org/10.1037/a0018124.
Gresham, F. M., Elliott, S. N., Vance, M. J., & Cook, C. R. (2011). Comparability of social skills rating system to the social skills improvement system: Content and psychometric comparisons across elementary and secondary age levels. *School Psychology Quarterly*, *26*, 27–44. http://dx.doi.org/10.1037/a0022662.
Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer.
Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. http://dx.doi.org/10.1080/10705519909540118.
Kang, T., & Chen, T. (2008). Performance of the generalized S-item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*, 391–406. http://dx.doi.org/10.1111/j.1745-3984.2008.00071.x.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum Associates.
MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130–149. http://dx.doi.org/10.1037/1082-989X.1.2.130.
Malecki, C. K., & Elliott, S. N. (2002). Children's social behaviors as predictors of academic achievement: A longitudinal analysis. *School Psychology Quarterly*, *17*, 1–23.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.
Matson, J. L., & Boisjoli, J. (2007). Differential diagnosis of PDD-NOS in children. *Research in Autism Spectrum Disorders*, *1*, 75–84. http://dx.doi.org/10.1016/j.rasd.2006.09.001.
McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, *9*, 49–57. http://dx.doi.org/10.1177/014662168500900105.
Merrell, K. W. (2002). *School social behavior scales, second edition.* Eugene, OR: Assessment-Intervention Resources.
Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176. http://dx.doi.org/10.1177/014662169201600206.
Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50–64. http://dx.doi.org/10.1177/01466216000241003.
Orlando, M., & Thissen, D. (2003). Further investigation of the performance of $S-\chi^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*, 289–298. http://dx.doi.org/10.1177/0146621603027004004.
Ostini, R., & Nering, M. L. (Eds.). (2006). *Polytomous item response theory models*. Thousand Oaks, CA: SAGE Publications, Inc. http://dx.doi.org/10.4135/9781412985413.
Parker, J. G., & Asher, S. R. (1987). Peer relations and later personal adjustment: Are low-accepted children at risk? *Psychological Bulletin*, *102*, 357–389. http://dx.doi.org/10.1037/0033-2909.102.3.357.
Raykov, T., & Marcoulides, G. A. (2015). Intraclass correlation coefficients in hierarchical design studies with discrete response variables: A note on a direct interval estimation procedure. *Educational and Psychological Measurement*, *75*, 1063–1070. http://dx.doi.org/10.1177/0013164414564052.
Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., ... Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks. *Medical Care*, *45*, 22–31.
Rennaissance Learning, Inc (2007). *Understanding STAR assessments.* Wisconsin Rapinds, WI: Renaissance Learning, Inc.
Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior assessment system for children* (2nd ed.). Circle Pines, MN: American Guidance Service.
Roff, M., Sell, B., & Golden, M. M. (1972). *Social adjustment and personality development in children.* Minneapolis, MN: University of Minnesota Press.
Salvia, S., Ysseldyke, J., & Bolt, S. (2010). *Assessment: In special and inclusive education.* Belmont, CA: Wadsworth Publishing.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monogrpaph Supplement*, *No. 17*. http://dx.doi.org/10.1007/BF02290599.
Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464. http://dx.doi.org/10.1214/aos/1176344136.
Sparrow, S. S., Cicchetti, D., & Balla, D. A. (2005). *Vineland adaptive behavior scales-2nd edition manual.* Minneapolis, MN: NCS Pearson, Inc.
Stapleton, L. M. (2013). Using multilevel structural equation modeling techniques with complex sample data. In G. R. Hancock, & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 521–562) (2nd ed.). Charlotte, NC: Information Age Publishing.
Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment*, *80*, 217–222.
Suen, H. K. (2008). Measurement. In N. Salkind (Ed.), *Encyclopedia of educational psychology*. Thousand Oaks, CA: Sage.
Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence*, *34*, 120–151. http://dx.doi.org/10.1177/0272431613511332.
U.S. Department of Education: National Center on Response to Intervention (2010t). *Review of progress-monitoring tools [review of STAR Math].* Washington, DC: Author.
Walker, H. M., & McConnell, S. (1995). *Walker–McConnell scale of social competence and school adjustment, elementary version.* San Diego, CA: Singular.
Wentzel, K. R. (1993). Does being good make the grade? Social behavior and academic competence in middle school. *Journal of Educational Psychology*, *85*, 357–364. http://dx.doi.org/10.1037/0022-0663.85.2.357.
Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245–262. http://dx.doi.org/10.1177/014662168100500212.
Zins, J. E., Weissberg, R. P., Wang, M. C., & Walberg, H. J. (Eds.). (2004). *Building academic success on social and emotional learning: What does the research say?*. New York: Teachers College Press.