

2014

VALIDATION OF A SOCIAL SKILLS CONSTRUCT USING MULTITRAIT MULTIMETHOD AND GENERALIZABILITY APPROACHES

Monica Mabe
University of Rhode Island, monbeebe@gmail.com

Follow this and additional works at: https://digitalcommons.uri.edu/oa_diss

Terms of Use

All rights reserved under copyright.

Recommended Citation

Mabe, Monica, "VALIDATION OF A SOCIAL SKILLS CONSTRUCT USING MULTITRAIT MULTIMETHOD AND GENERALIZABILITY APPROACHES" (2014). *Open Access Dissertations*. Paper 240.
https://digitalcommons.uri.edu/oa_diss/240

This Dissertation is brought to you by the University of Rhode Island. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons-group@uri.edu. For permission to reuse copyrighted content, contact the author directly.

VALIDATION OF A SOCIAL SKILLS CONSTRUCT
USING MULTITRAIT MULTIMETHOD
AND GENERALIZABILITY APPROACHES

BY
MONICA MABE

A DISSERTATION SUBMITTED IN PARTIAL FULLFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
SCHOOL PSYCHOLOGY

UNIVERSITY OF RHODE ISLAND

2014

DOCTOR OF PHILOSOPHY DISSERTATION

OF

MONICA MABE

APPROVED:

Dissertation Committee:

Major Professor

W. Grant Willis

John Stevenson

Susan Brand

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2014

Abstract

Social skills are important components of social-emotional functioning that allow children to be successful in school, both socially and academically. A single, agreed-upon definition of social skills has not been identified in the literature, however, which has led to variations in the measurement and assessment of social skills. Issues of measurement may be linked to the ineffectiveness of school-based social-skills interventions. Commonly used conceptualizations and operationalizations of social skills are presented, as well as a review of issues surrounding social-skills interventions. The use of a multitrait-multimethod approach is presented for the use of establishing a unified set of social skills and the use of generalizability theory is examined as a psychometrically-based approach to developing a measure for observing social skills. The assessed skills included six social skills, drawn from the Social Skills Improvement System – Rating Scales, and two academic skills. Skills were assessed using behavioral observation and rating scales. Convergent and discriminant validity was evaluated for a social skills construct, or a homogenous set of social skills. The reliability of the observational strategy was evaluated in order to assess the optimal number of occasions and skills needed in order to obtain adequate degrees of reliability. Results indicated that a homogenous set of skills could be identified and that a behavioral observation strategy could be used reliably to assess social skills. Results are discussed in terms of applied use for the measurement strategy in school settings for formative assessment and in terms of directions for future research.

Acknowledgement

I would like to express my thanks and appreciation to those who helped me accomplish this goal and complete my dissertation. First, I would like to thank my major professor, W. Grant Willis, for his guidance and support throughout my graduate career. Second, my appreciation goes out to the students and staff at the participating school for their willingness to assist me in my research. Next, I would like to thank my friends and family for their love and encouragement when I needed it most. Finally, and most significantly, I would like to thank my husband for his endless and tireless support, for never letting me give up or settle, for pushing me forward, and for being a constant source of encouragement.

Table of Contents

	page
Abstract	ii
Acknowledgments	iii
Table of Contents	iv
List of Tables	vii
List of Figures	viii
Chapter I: Introduction	1
Statement of the Problem	1
Critical Review of the Literature	3
Social Skills Background	3
Definitions and Conceptualizations	3
Operationalization and Assessment	4
Social Skills Interventions	5
Effectiveness	5
Progress Monitoring and Direct Observation	8
Measurement Approaches to Establishing Validity	10
Multitrait-Multimethod	10
Generalizability Theory	11
Purpose of the Study	12
Chapter II: Method	14
Participants and Setting	14
Design	17

Table of Contents (Continued)

	page
Measures	18
Dependent Variables	20
Procedures	22
Informed Consent/Assent	22
Training Procedures	24
Direct Observation	25
Rating Scales	27
Chapter III: Results	28
Multitrait-Multimethod Analyses	28
Social Skills Domain	32
Convergent Validity for Social Skills	33
Academic Skills Domain	33
Homomethod Discriminant Validity	33
Heteromethod Discriminant Validity	34
Test-Retest Reliability	34
Generalizability and Decision Studies	35
Interobserver Agreement	38
Descriptive Analyses	39
Chapter IV: Discussion	41
Psychometric Findings	42
Multitrait-Multimethod Study	43

Table of Contents (Continued)

	page
Social Skills Domain	43
Convergent Validity for Social Skills	44
Academic Skills Domain	44
Homomethod Discriminant Validity	45
Heteromethod Discriminant Validity	45
Test-Retest Reliability	46
Generalizability Study	47
D-Study	49
Interobserver Agreement	50
Cultural Considerations	51
Implications	51
Limitations	54
Future Directions	56
Summary and Conclusions	57
Appendix A: Expanded Multitrait-Multimethod Matrix	60
Appendix B: Parent/Caregiver Consent Form	62
Appendix C: Student Assent Form	66
Appendix D: Teacher Consent Form	69
Literature Cited	72

List of Tables

	page
Table 1. Sample Characteristics	15
Table 2. Observation Matrix	26
Table 3. Teacher Rating Correlations	28
Table 4. Teacher Rating and Behavioral Observation Correlations	29
Table 5. Teacher and Student Rating Correlations	29
Table 6. Behavioral Observation Correlations	30
Table 7. Behavioral Observation and Student Rating Correlations	30
Table 8. Student Rating Correlations	31
Table 9. Condensed MTMM Matrix	32
Table 10. Teacher Test-Retest Reliability	34
Table 11. G-Study Results	36
Table 12. D-Study Results	37
Table 13. Interobserver Agreement Kappa Values	39
Table 14. Results of MANOVAs for Demographic Factors	40
Table 15. Median Multitrait-Multimethod Correlation Coefficients for Social Skills	46

List of Figures

	page
Figure 1. Percents of Variance Explained	36
Figure 2. D-Study Results for 1 Observer	38

VALIDATION OF A SOCIAL SKILLS CONSTRUCT USING
MULTITRAIT MULTIMETHOD AND GENERALIZABILITY APPROACHES

Chapter I: Introduction

Statement of the Problem

Social functioning plays an important role in students' abilities to thrive in school (Cappadocia & Weiss, 2011). Students with poor social skills are at-risk for internalized and externalized behavioral difficulties as well as poor academic achievement (Cook, Gresham, Kern, Barreras, Thornton, & Crews, 2008). Social-skills interventions occur in schools as a means of improving the social-emotional functioning of students with difficulties. Unfortunately, however, social-skills interventions often are ineffective (Cappadocia & Weiss, 2011; Gresham, Sugai, & Horner, 2001), and this may be related to associated assessment and measurement issues (Matson & Wilkins, 2009).

Proper assessment of skill deficits and skill performance in multiple settings may improve the overall effectiveness of social-skills interventions. The most common social-skills assessment methods are direct behavioral observation and behavioral rating scales (Matson & Wilkins, 2009). There are no current, agreed-upon criteria for defining social skills, however, and, as such, measurement strategies vary and focus on identifying specific behaviors to target for treatment (Matson & Wilkins, 2009); rating scales often vary greatly in the kinds of behaviors and skills they assess. Additionally, measures for monitoring the progress of social-skills interventions, which would improve differentiation of instruction and

potentially increase effectiveness, are not currently available (Gresham et al., 2010), although there are a few in development (Cummings, Kaminski, & Merrell, 2008; Stichter, Herzog, O'Connor, & Schimdt, 2012).

The current research aimed to use a multitrait-multimethod approach to evaluate the validity of a social skills construct. Additionally generalizability theory was employed to evaluate the psychometric properties of a behavioral measurement tool and its utility for using multiple observers in multiple settings to observe social skills. It was hoped that a reliable set of social skills could be identified and a useful method for progress monitoring of these skills could be established.

Critical Review of Literature

The following critical review focuses on theoretical issues surrounding the conceptualization of social-skills constructs and their measurement, aspects affecting the outcome, or overall effectiveness, of social-skills interventions, and measurement approaches to establishing reliability and validity.

Social Skills Background

Definitions and conceptualizations. Social skills may be conceptualized as a set of competencies that allow an individual to initiate and to maintain social relationships that contribute to peer acceptance and school adjustment (Luiselli, McCarty, Coniglio, Zorilla-Ramirez, & Putnam, 2005). Social skills also have been conceptualized as social cooperation skills that lead to successful school adjustment and positive peer relationships (Cummings et al., 2008). Crowe, Beauchamp, Catroppa, and Anderson (2011) suggest that appropriate social functioning is important across the lifespan and is the basis for forming lasting relationships. Gresham et al. (2001, pp. 333) conceptualized social skills as “specific behaviors that an individual uses to perform competently or successfully on particular social tasks.”

How social skills are defined or conceptualized has an impact on the manner in which they are operationalized, or measured. The operationalization and assessment of social skills, in turn, may impact the skills that are targeted for intervention as well as the intervention approach and its overall effectiveness in improving said skills.

Operationalization and assessment. As mentioned previously, no single, agreed-upon conceptualization or definition of social skills exists. As such, multiple social-skills constructs have been suggested as well as multiple methods of assessing said constructs. For the purposes of measurement, social skills are often broken down into distinct observable behaviors. In other words, social skills may be operationalized as the target behavior(s) to be observed, or measured. Most frequently, these distinct behaviors are measured through contrived social role plays or standardized rating scales (Matson & Wilkins, 2009). Matson and Wilkins's review in 2009 of social assessments found 40 questionnaire-style, or rating-scale, assessments and 8 role-play tests. More recently, Crowe et al. (2011) conducted a comprehensive review of assessment tools for social functioning in children and adolescents and found a total of 86 measures. In addition to standardized rating scales and contrived social role plays, the identified assessment styles included transcription and scoring of verbal responses to hypothetical situations, responding to video vignettes, self-monitoring, responding to computer-presented scenarios, and direct observation. Consistent with Matson and Wilkins's findings, rating scales were the most commonly used form of assessment.

Gresham (2002, as cited in Luiselli et al., 2005) suggested that a problem-solving model should be used to assess social skills; and that problem identification and problem analysis form the first two stages of this model and should incorporate rating scales and direct observations in tandem. In this manner, information may be obtained regarding the context in which social difficulty occurs and to what degree (i.e., compared to same-aged peers). This information then can be used in the

second two stages of the problem-solving process, namely plan, or intervention, implementation, and treatment evaluation. In this problem-solving model, rating scales are suggested to help identify target behaviors for intervention but may not serve as the best means of evaluating intervention effectiveness, or overall outcomes.

Social-Skills Interventions

Social skills play an important role in child development, in general, and play a pivotal role in a student's ability to thrive in a school environment (Capadocia & Weiss, 2011). Luiselli et al. (2005) suggested that success in school depends as much on social skills as academic performance, namely because effective social skills help children to form friendships, respond appropriately to classroom expectations, and build positive relationships. Well-developed social skills are linked to positive social and school-related outcomes whereas poorly developed social skills may place students at risk for poor social and school-related outcomes (Cook et al., 2008). Gresham, Cook, Crews, and Kern (2004) have suggested that social skills may serve as academic enablers because they appear to help facilitate academic performance and often are linked with academic achievement.

Students who lack social-emotional skills often receive school-based interventions in order to remediate associated difficulties and to facilitate social-emotional development. These interventions, however, often unfortunately are ineffective.

Effectiveness. Within the last three decades, multiple meta-analyses have been conducted assessing the effectiveness of social-skills interventions (see Mabe,

2013 for a more detailed examination). Reviews of the meta-analytic literature (Cook et al., 2008; Gresham et al., 2004; Gresham et al., 2001) revealed two major findings: (a) many studies used outcome measures that were not directly linked to the skills taught during intervention, and (b) the majority of studies relied solely on outcome-based evaluation rather than formative assessment throughout.

Standardized rating scales, discussed previously, are an example of outcome-based measures that are frequently used to assess the effectiveness of social-skills interventions. When considering the first major finding, it is important to note that rating scales are often composed of hundreds of items. It is easy to see how these types of measures can assess a large number of skills and not necessarily focus on the skills targeted within an intervention. Additionally, in consideration of the second major finding, these types of measures provide summative information in a global manner that may not indicate a student's standing on specific skill components.

Rating scales are typically used as a general outcome measure (GOM), which do not provide information about specific skill deficits and simply provide an overall general description of skills (Hosp, Hosp, & Howell, 2007). Formative assessment, however, utilizes change-sensitive measures that allow one to observe small changes in performance over brief periods of time (Burns & Coolong-Chaffin, 2006). This method of assessment allows for a more detailed examination of a student's performance on specific skills over short periods of time and may be beneficial for instructional planning. A GOM may be used to determine if a student can perform particular tasks subsequent to an intervention. A formative assessment

model differs from a GOM because it is change-sensitive and uses multiple measurements of skills throughout an intervention to detect performance changes, which inform instructional decisions throughout (Hintze, Christ, & Methe, 2006). Relying on an outcome measure limits the individualization of an intervention that may be necessary to improve its effectiveness.

Overreliance on outcome-based measurement and poor individualization of instruction are major contributing factors to the ineffectiveness of social-skills interventions. Linked to these issues, and also contributing to ineffectiveness, are the issues of: (a) conceptual clarity, (b) instructional format, and (c) generalization of instruction. First, as previously discussed, there are many varied conceptualizations of social skills. Behaviors may be targeted for intervention based on assessments with a specific conceptualization (Matson & Wilkins, 2009) and omit potential skills not included in that particular conceptualization. In other words, skills that may benefit from instructional support may not be included in interventions because they were not initially included as part of the identified social-skills construct. Skills that often are grouped (i.e., taught and measured) together may not be homogenous and may attenuate the reliability of measurement (Mabe, 2013). Second, the instructional format of social-skills interventions often focuses on acquisition deficits (have not learned skill) rather than performance deficits (do not perform a previously learned skill) and does not tailor instruction to account for individual differences (Gresham et al., 2001). Interventions may teach a number of skills to a group of students, all of who may experience different skill deficits. A failure to differentiate instruction to meet specific needs of students

within a group may make the intervention less effective. Finally, interventions often lack generalization instruction, resulting in a failure of instruction to generalize to new situations (Cappadocia & Weiss, 2011; Gresham et al., 2001). Here, students may be instructed and learn to perform skills in one setting (e.g., instructional room outside of the classroom) but not in another (e.g., classroom or playground). Issues of assessment are found among each of the discussed issues affecting intervention effectiveness, specifically, concerning conceptual clarity of social skills assessment and formative assessment approaches.

Progress monitoring and direct observation. Progress monitoring is a formative assessment approach used in conjunction with a multi-tiered format of intervention in schools called response-to-intervention (RTI). The RTI process is a complex one and depends on valid, easily administered, brief, change-sensitive measures to inform interventionists about student progress on specific skills in order to make decisions regarding their progress (Burns & Coolong-Chaffin, 2006; Hosp et al., 2007). These measures need to have a high level of reliability in order to make accurate instructional decisions from obtained data. Progress monitoring provides the means of evaluating instruction and decision making regarding instructional modifications (Fletcher & Vaughn, 2009; Stecker, Lembke, & Foegen, 2008). In other words, progress-monitoring tools are essential to effective interventions because they provide data for decision making about student needs and differentiation of instruction.

Social skills are often challenging to assess in a manner that is reliable, generalizable, and efficient (Cummings et al., 2008). Recently, some measures have

begun to be developed in an effort to produce progress-monitoring tools for social-skills interventions that meet these criteria. Brief rating scales have been developed (Gresham & Elliott, 2008; Gresham et al., 2010) but remain in the style of traditional rating scales and may not be adequate for progress-monitoring purposes.

Cummings et al. (2008) have begun developing the Initiation-Response Assessment (IRA) Code and Stichter et al. (2012) have begun developing the General Social Outcome Measure (GSOM). Both assessment tools are performance-based assessments that use behavioral observation during analog, or contrived, social situations and are meant to be administered multiple times throughout an intervention. The IRA uses a behavioral code to assess the frequency and quality of social interactions. The GSOM uses a 5-point likert scale to rate accuracy and social appropriateness of behavioral responses during analog situations. Behavioral observation appears to be a preferred method for the development of progress monitoring tools for social skills.

Although behavioral observations may be well suited to progress monitoring, Hintze and Matthews (2004) found that direct behavioral observation often has low reliability even when interrater agreement is high. They suggested that direct observation may not be as reliable of a method as it is often believed to be in the field of school psychology. In another study that assessed the reliability of a behavioral observation tool (Mabe, 2013), three occasions of observation were found to be inadequate and a minimum of five occasions would be required for accurate assessment of skills. In the same study, the skills that were assessed did not cluster well together as hypothesized, which attenuated reliability. In order to

obtain an adequate level of reliability for making instructional decisions, it appears that one would need to assess a validated homogenous group of skills on a sufficient number of occasions for a representative sampling of skill performance. Four occasions of observation may prove adequate for skill assessment if a higher level of interobserver agreement and skill homogeneity could be obtained.

Measurement Approaches to Establishing Validity

Multitrait-multimethod. A multitrait-multimethod approach is a method of assessing validity for a construct by evaluating convergent and discriminant validity simultaneously. Campbell and Fiske (1959) explained that convergent validity is necessary to validate a trait (e.g., social skills) but that discriminant validity is also required. In other words, one must be able to establish what a trait is, as well as what it is not. To do this, more than one trait and more than one method must be used during the validation process; the traits should be theoretically unrelated in order to ensure that discriminant validity can be obtained. For example, social skills and academic skills could be assessed using rating scales and behavioral observation because the skills and methods are theoretically independent of each other. Although the link between social and academic skills has long been studied and it is widely recognized that one may often have an impact on the other (Welsh, Parke, Widaman, & O'Neil, 2001), they are viewed as separate and distinct constructs; certainly, the assessment of one could not be substituted for the other with any degree of accuracy.

After identifying at least two theoretically independent traits, correlations can then be calculated between each trait as measured by each method in order to

evaluate reliability and validity. Here, the index of reliability should exceed convergent validity values, and convergent validity values should exceed discriminant validity values (see Campbell & Fiske, 1959). Values that do not follow this order are an indication of poor validity for the methods or traits. For example, in the current study, it was anticipated that social and academic skills may have some degree of correlation, as one may be predictive of the other in some instances (Welsh et al., 2001). A high level of correlation between these traits, however, would indicate poor validity of trait constructs and would indicate that the two traits were theoretically linked in some way.

Generalizability theory. Generalizability theory (GT) is an extension of classical test theory (CTT) that can be used for assessing the reliability of behavioral measures. CTT is a major perspective in psychometric assessment that is used to evaluate measurement variability. In CTT, variability in test-scores is partitioned into two areas: (a) variance due to true scores, and (b) variance due to error. The major assumption in CTT is that error is randomly distributed and comes from sources unrelated to true differences in the assessed trait. GT extends CTT in a number of ways: (a) recognizing multiple sources of measurement error, (b) estimating each source of measurement error separately, (c) indexing the magnitude of each source of error, (d) distinguishing between relative (i.e., normative or inter-individual) and absolute (i.e., ipsative or within-individual) decisions, and (e) differentiating between generalizability and decision studies (Shavelson, Webb, & Rowley, 1989). Of particular interest for the purposes of this study is the use of GT to account for multiple sources of error, estimate the

magnitude of error for each source, and differentiate between generalizability and decision studies. Reliability estimates from GT studies account for expected error as well as additional error sources, which are important for the evaluation of behavioral measures (Hintze & Matthews, 2004). For example, in GT, one can account for error attributed to multiple observers and multiple settings when using behavioral observation.

As noted, GT differentiates between two phases of a study: (a) generalizability studies and (b) decision studies. These two phases work together to estimate an optimal level of reliability for a measure. The generalizability-study phase results in a *g*-coefficient that provides an estimate of reliability for a measure and can be interpreted similarly to the *r*-coefficient in CTT; the Generalizability study also estimates the magnitude of each source of error. The decision-study phase uses data from the generalizability study to estimate the impact that changes to a measurement strategy can have in order to minimize error (Shavelson et al., 1989). In other words, one can estimate how adjustments to sources of error may influence reliability. For example, one could estimate how many observations need to be conducted of a particular skill in order to obtain a particular level of reliability.

Purpose of the Study

There is no single, agreed-upon definition or conceptualization of social skills in the literature. Social skills are often operationalized based on specific observable behaviors that differ across assessments. A number of assessments exist that assess various conceptualizations of social skills. The current study to uses a multitrait-multimethod approach to evaluate the validity of a social-skills construct.

Additionally, an effective change-sensitive progress-monitoring tool is necessary to help social-skills instructors gauge student progress on specific skills, differentiate instruction appropriately, and improve the overall effectiveness of social-skills interventions. The current study uses G theory to develop an observational, formative assessment tool for social-skills interventions that could be used for progress monitoring and decision-making purposes. It is hoped that the identification of a unitary set of social skills and the establishment of a method to monitor progress reliably will be able to improve program effectiveness and student outcomes.

Chapter II: Method

Participants and Setting

Participants were 20 middle and high-school students from specialized classrooms in sixth through twelfth grade (ages 12 to 18 yrs., $M = 15$ yrs., $SD = 2.1$ yrs., Median = 15 yrs.) attending an alternative public day school run by an educational collaborative in the northeastern part of the United States. Sample size was selected given considerations for the statistical analyses that were conducted and with consideration of the limited size of the population and the lengthy process required to obtain consent. All grade levels of the participating school were represented due to the limited student population from which to draw a sample. The participating school is comprised of two specialized middle-school classrooms and three specialized high-school classrooms run by an educational collaborative. All students meet Diagnostic and Statistical Manual of Mental Disorders (DSM IV, TR) diagnostic criteria for one or more mental-health conditions. Students attending the collaborative school are in attendance because they have not been able to be successful in a traditional educational setting and have been referred to the school by their member district (i.e., school district that participates with the collaborative). Students at the collaborative school often experience severe social-emotional and/or behavioral difficulties that require educational modifications in order for them to be successful. Students may have developmental delays or cognitive impairments as well. Students are referred to the collaborative school from member districts in the surrounding area; the student population is diverse and consists of students from urban, low socio-economic status (SES) areas as well as suburban middle-class areas. SES was estimated by participation in the school's lunch program: Students who qualified for free lunch

were estimated to come from families of low SES, those who were eligible for a reduced-price lunch were estimated to come from families of middle to low SES, and those who paid the full price for lunch were estimated to come from families of middle to above SES. The majority of the sample (i.e., 75%) qualified for free lunch (low SES), 0% for reduced-price lunch (medium to low SES), and 25% paid the full price for lunch (medium to above SES). The characteristics of the sample are represented in Table 1.

Table 1. *Sample Characteristics*

Characteristic		<i>n</i>	%
Gender	Male	17	85%
	Female	3	15%
Ethnicity	White	15	75%
	African American	4	20%
	Hispanic	1	5%
SES	Low	15	75%
	Low to middle	0	0%
	Middle to above	5	25%
Age	11-12	3	15%
	13-14	6	30%
	15-16	6	30%
	17-18	5	25%
Grade level	Sixth	3	15%
	Seventh	3	15%
	Eighth	3	15%
	Ninth	2	10%
	Tenth	3	15%
	Eleventh	1	5%
	Twelfth	4	20%

Characteristic		<i>n</i>	%
	Mood Disorder (NOS, Major Depressive, Bipolar)	11	55%
	Attention Deficit Hyperactivity Disorder	9	45%
	Autism Spectrum Disorder	6	30%
Characteristic		<i>n</i>	%
Diagnosis	Post Traumatic Stress Disorder	4	20%
	Anxiety Disorder	4	20%
	Oppositional Defiant Disorder	3	15%
	Reactive Attachment Disorder	1	5%
	Obsessive Compulsive Disorder	1	5%
	Conversion Disorder	1	5%

Several considerations should be made regarding the contribution of cultural influences shared by the sampled population to the expression of social skills. First, socioeconomic status indirectly may be linked to poor social skills. Children from low SES backgrounds often qualify for free or reduced meals at school in order to reduce academic and behavioral difficulties due to hunger (as mentioned previously, the number of students who qualified for free or reduced lunch was used as an indicator of SES). Jyoti, Frongillo, and Jones (2005) found that food insecurity over time is related to decline in reading and math test performance, increase in weight, and impairment of social skills. Results from that study also indicated that children from low SES backgrounds may experience difficulties over time in social and academic areas if their basic dietary needs are not met.

Additionally, the sample population is clinical in nature and the participants presented with marked social difficulties as a result. It is commonly understood that

children and adolescents with Autism Spectrum Disorder frequently exhibit social interaction and communication difficulties. Other disorders found within the sample population, such as Attention-deficit Hyperactivity Disorder, Mood Disorder, Anxiety Disorder, and Oppositional Defiant Disorder, are also associated with social difficulties, although they may originate for different reasons. It is important to note that the sample population is distinctly different from a typical population of students in a regular-education setting.

Design

Two broadly defined traits and three methods were selected for the multitrait-multimethod design. The first trait, social skills, was more narrowly delineated into six social domains (representing the construct of interest for validation). The second trait, academic skills, was more narrowly delineated into two academic domains, that is reading and math, and was selected as a discriminative construct. Academic skills were selected as the discriminant trait in this study because they are theoretically separate from social skills and can be measured as such. Although studies such as those conducted by Arnold, Kupersmidt, Voegler-Lee, and Marshall (2012) and Demaray and Jenkins (2011) highlight the relationships that exist between social and academic skills, these studies also demonstrate that these two constructs are disparate ones and can be measured separately. Each construct was assessed with three methods: (a) teacher rating scale, (b) student rating scale, and (c) behavioral observation. The dependent variables were the obtained scores for skill performance for each social-skill observation (averaged across 4 occasions), obtained scores for academic and social rating scales (averaged across 2

occasions) and academic skill observation (obtained from performance scores on a standardized test). The multitrait-multimethod matrix is illustrated in Appendix A.

The generalizability study was conceptualized as a three-facet, partially nested design with occasions (4 levels) and skills (6 levels) as crossed facets, and students (20) nested within observers (3 levels). For practical reasons, students were nested within observers because each observer was randomly assigned to particular students; in order for them to be completely crossed (i.e., independent), each observer would have had to observe every student on all skills and occasions, and this was not feasible for the current study. The dependent variable was the observational outcome as a percentage of intervals observed for each skill for each student ($N = 480$ measures). This design allowed for an estimation of variance components for (a) occasion; (b) skills; (c) observers; (d) students nested within observers; (e) the interactions between skills and occasions, skills and observers, occasions and observers, skills and students nested within observers, and occasions and students nested within observers; and (f) residual error.

Measures

Rating-scale data for social skills and academic skills were obtained using the Social Skills Improvement System (SSIS) Rating Scales (Gresham & Elliott, 2008). The SSIS is a nationally normed and standardized evaluation tool for assessing Social Skills, Competing Problem Behaviors (which was not used in the present study), and Academic Competence. The SSIS is based on factor analytic research that supports evidence for an overall score for social skills as well as subscales, or factors, of Communication, Cooperation, Assertion, Responsibility, Empathy, Engagement, and Self-Control. The subscales for Academic Competence are Reading Achievement, Math Achievement, and

Motivation to Learn. Subscales are also provided for Competing Problem Behaviors, but were not used in the present study. Behaviors for direct observation of social skills were derived from the Social Skills subscales on the SSIS.

Observational data for social skills was obtained using Metryx, a tool designed for tracking student progress on a number of skills. Metryx was developed by Stephanie Castilla and Shawn Rubin (COO and CEO of Metryx, respectively) at a charter school in the northeastern part of the United States to supplement traditional observation techniques for the purposes of progress monitoring and decision-making in social-skills interventions. Metryx uses iPad technology for recording behavioral observations and providing instant feedback about a student's performance in various skills. Metryx is also available for iPhone as well as Android phone and tablets and can be accessed via internet on a computer as needed.

Observational data of academic performance was obtained using standardized measures of achievement for reading and math. Reading and Math scores were obtained primarily from the Wechsler Individual Achievement Tests – Third Edition (WIAT-III; Wechsler, 2009). The WIAT-III is a nationally standardized assessment of academic achievement and contains multiple subscales. The WIAT-III math and fluency scales are of interest for the present study because they allow for a direct observation of academic skill performance, provide standardized scores for comparison, and assess the academic skills described in the SSIS Academic Competence subscale. Reading and math scores were obtained from existing student records as participants had all completed achievement testing as part of their educational planning. As stated, the majority of academic scores (e.g., reading and writing) came from the WIAT-III. Some students,

however, had completed the Kaufman Test of Educational Achievement – Second Edition (KTEA-II; Kaufman & Kaufman, 2004) or the Woodcock-Johnson Tests of Achievement – Third Edition (WJ-III; Woodcock, McGrew, & Mather, 2001) and academic scores were obtained from these measures. These assessments are also nationally standardized assessments of academic achievement that provide standardized scores for similar reading and writing tasks (Flanagan, Alfonso, & Dixon, 2014, pp. 38-41). Scores from the KTEA-II and WJ-III may be reasonably compared to the WIAT-III reading and writing scores given that each assessment utilizes standard scores of 100 with standard deviations of 15.

Dependent variables. The dependent variables for social skills were teacher and student ratings as reported on the SSIS Rating Scales, teacher and student forms, as well as observational ratings of successful completion of specified skills. Various skills were selected for observation from three subscales on the SSIS (Communication, Cooperation, and Engagement); skills representative of other subscales were not selected due to potential low availability to observe in typical classroom settings. Six skills were identified for observation (a) conversation, (b) nonverbal communication, (c) classroom participation, (d) follow expectations, (e) group participation, and (f) interaction.

“Conversation” was defined as using appropriate conversational skills such as responding when spoken to, using appropriate tone and volume, and taking turns while speaking.

“Nonverbal communication” was defined as using nonverbal communication appropriately during conversation such as making eye contact, facing appropriately, and maintaining appropriate distance between speakers. “Classroom participation” was defined as being actively or passively involved, as appropriate to situation, in instruction;

examples included volunteering to answer questions, taking notes, following along in text, engaging in class discussion, and working on assignments. “Follow expectations” was defined as engaging in appropriate classroom behavior, such as following directions, completing tasks without disrupting others, following classroom rules, and ignoring distractions. “Group participation” was defined as engaging in group interactions; examples included joining activities that have already started, participating in games or group activities, and inviting others to join in activities. Finally, “Interaction” was defined as the quality, or appropriateness, of social engagement such as ease of engagement and positive interactions. Consistent with the instructions for the SSIS, each of these items were rated by both teachers and students on a 4-point Likert scale with the letters “N,” “S,” “O,” and “A” assigned to the values 0, 1, 2, and 3 respectively. A response marked “N” indicated that the student never exhibits the behavior, “S” indicated that they seldom exhibits the behavior, “O” indicated that they often exhibit the behavior, and “A” indicated that they almost always exhibit the behavior.

The rubric for scoring the behavioral observations was as follows: Conversation, Follow Expectations, and Interaction were observed and rated using the slider tool for rubric scoring on Metryx. The slider tool allows the observer to rate a student’s performance of a skill on a scale of 0 to 100%, indicating the percentage of skill criteria completed. A rating of 1-20% indicates that a student was not very successful in the completion of the skill, 21-40% indicates they were somewhat successful, 41-60% indicates they succeeded in completing about half of the criteria, 61-80% indicates they were mostly successful, and 81-100% indicates they completed criteria nearly flawlessly. Nonverbal Communication, Classroom Participation, and Group Participation were

observed using momentary time sampling for 15-minute observations with 30-second intervals. In this manner, a student was observed at the end of each 30-second interval and rated as engaging in the behavior or not. This type of observation resulted in a percentage from 1 to 100 indicating the percentage of intervals in which the student was successfully engaged in the observed behavior.

The dependent measures for academic skills were teacher ratings as reported on the SSIS, student ratings as reported on a similar form, and performance scores from standardized achievement assessments as reported in student records. Teachers and students rated academic performance on a 5-point Likert scale comparing student performance to that of other students. Academic skills were rated as being in the Lowest 10%, Next Lowest 10%, Middle 40%, Next Highest 20%, or Highest 10% with the values 1, 2, 3, 4, and 5 respectively. The student SSIS form does not include an academic skills section, however, a form was developed (Appendix A) by the researcher, based on the teacher SSIS Academic Competence scale, for students to complete. Reading and math performance scores were obtained from the WIAT-III ($n = 11$), KTEA-II ($n = 5$), and WJ-III ($n = 4$). Students completed tasks that involved reading aloud for a timed period and performing math computations. Reading and math tasks were similar across assessments.

Procedures

Informed Consent/Assent

Informed consent was obtained from parents/caregivers for participation in the study for students under the age of 18, and assent was obtained from all students. Students who were 18 years of age provided informed assent for participation and a

follow-up phone call was provided to the parent/caregiver by the student's school-based clinician to notify them of their child's participation in the study and to clarify any questions the parent/caregiver might have. All parents/caregivers were provided with a phone call from their child's school-based clinician to explain the study and answer questions prior to the consent form being sent home. Students received "credits," or points to be used toward school-based rewards upon return of the parent/caregiver form. Students received credits regardless of whether or not the parent/caregiver agreed to student participation in the study. Parent/caregiver consent forms (Appendix B) were sent home and returned to school with the student. Consent forms were sent home with 22 students and 86% (19 parents) were signed and returned. Two of the 22 parents/caregivers declined to have their child participate in the study. Student assent forms (Appendix C) were explained by the researcher to each student individually upon obtaining parent/caregiver consent. Students who were 18 years of age ($n = 3$) completed the assent process as previously outlined. Teacher consent (Appendix D) was obtained at a staff meeting after the research study was discussed and all questions answered. Rating scales were not completed by teachers who did not sign consent.

Parent/caregiver consent and student assent were the primary inclusion criteria. Students also needed to reside outside of a group home setting in order to simplify the consent process as well as spend the majority of their day in the classroom. Due to the nature of the school population, many students have difficulty staying in class for extended periods of time and often walk the halls or take movement breaks in the gym. In order to be included in the study, participants needed to be located in the classroom reliably for an extended period of time on most days. All students for whom

parent/caregiver consent was obtained assented to participate in the study. Data were not collected from students within each classroom for whom informed consent and assent were not obtained. All students were treated in a manner consistent with the ethical guidelines of the American Psychological Association, the National Association of School Psychologists, and the Institutional Review Board of the University of Rhode Island.

Training Procedures

Two females enrolled in a psychology undergraduate program served as observers, in addition to the researcher, for course credit. Observers were trained in direct observational methods, the use of Metryx, and in how to navigate the classroom settings appropriately (e.g., consideration of student location within the classroom, appropriate classroom demeanor, how to use timing devices properly during observational periods, consideration of skill presentation in a clinical population). Assistants attended four hours of training (divided into two sessions) conducted by the researcher. During the training sessions, assistants discussed operational definitions of the behaviors to be observed and were trained in the observation methods used in the present study (e.g., momentary time sampling and rubric scoring). Additionally, assistants practiced observation skills while observing video recordings of children in classroom settings. Assistants practiced observations on video recordings until 80% agreement was obtained between the assistants and the researcher.

All assistants were required to provide documentation of education and training in the “Responsible Conduct of Research” and of an official criminal background check

prior to conducting observations in the school. All rating scales were hand-scored by the researcher and double-scored on a separate occasion.

Direct Observation

Students were observed for social-skill performance during the naturally occurring day in the classroom and transition periods such as breakfast, lunch and free time. Dates, times, and subject matter being studied during an observation were recorded. The skills “Conversation” and Nonverbal communication” were observed on the same occasions as one skill pair representing the Communication subscale. “Classroom participation” and “Follow expectations” were observed on the same occasions as a second skill pair representing the Cooperation subscale. Finally, “Group participation” and “Interaction” were observed on the same occasions as a third skill pair representing the Engagement subscale. Observers were randomly assigned to students. Each participating student was observed on four occasions on each skill pair; each observation was 15 minutes long, divided into thirty-second intervals. Every participating student was observed for 15 minutes in a classroom setting on 12 separate occasions for a total of 180 minutes (3 hours).

Social-skill observations were conducted using momentary time sampling and rubric scoring. Each skill pair included one skill to be observed using momentary time-sampling (Conversation, Class participation, Group participation) and one skill using rubric scoring (Nonverbal communication, Follow expectations, Interaction). The momentary time sampling procedure required observers to observe students across a 15-minute period, whereas the rubric scoring system required observers to evaluate the percentage of skill criteria completed during the 15-minute period. Thus, the researcher

paired them together in order to maximize the productivity of time spent in observation.

Table 2 illustrates the social-skill observational matrix for the present study.

Table 2: *Observation Matrix*

		Skill					
		Pair 1		Pair 2		Pair 3	
		Nonverbal	Class	Follow	Group		
Observer	Occasion	Conversation	Comm.	Part.	Expect.	Part.	Interaction
A	1	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}
	2	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}
	3	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}
	4	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}
	5	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}
	6	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}	S_{1-8}
B	1	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}
	2	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}
	3	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}
	4	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}
	5	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}
	6	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}	S_{9-12}
C	1	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}
	2	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}
	3	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}
	4	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}
	5	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}
	6	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}	S_{13-20}

The Undergraduate assistants spent between 4 and 18 hours each week observing the participating students on the specified social skills. Observations were collected for 3 months from December 2013 through February of 2014 (with 7 days off for a school

break in December, 1 day off for professional development, and 3 days off for school cancellations due to snow) until all observations were completed.

Additionally, inter-rater observations were conducted for each of the students on the fourth occasion for each skill. Secondary observers randomly were assigned to students and they conducted an observation on each skill simultaneously with the primary observer for each student. This provided inter-rater information for each of the 20 students on one occasion for each skill.

Rating scales. Academic and social skills were assessed using the SSIS and the companion academic form for students, developed by the researcher, on 2 occasions (2 to 3 weeks apart) in order to obtain test-retest reliability. Participating teachers were asked to complete the SSIS and return it to the researcher. Participating students were asked to complete the SSIS with their school-based clinician or the researcher so that any questions they had while completing the form could be answered. School-based clinicians were asked to return completed student forms to the researcher. The SSIS consists of 83 items to be answered on a Likert-type scale as previously described. Participants were only asked to complete the social skills and academic sections of the SSIS, which consists of 53 items. The SSIS forms take approximately 10-15 minutes to complete. Thus, participants spent approximately 20-30 minutes completing rating scales over the two testing periods.

Chapter III: Results

Six sets of analyses were conducted: (a) Multitrait-multimethod analyses, (b) Test-retest reliability analyses, (c) G-study analyses, (d) D-study analyses, (e) Kappa analyses of interobserver agreement, and (f) MANOVA analyses of demographic characteristics.

Multitrait-Multimethod Analyses

Pearson product-moment correlations were calculated to assess convergent and discriminant validity for a social skills construct. Scores were averaged across ratings or observational occasions for comparison. Tables 3 through 8 illustrate the obtained correlations for the multitrait-multimethod matrix. Correlations marked with an asterisk (*) are significant at the .05 alpha level.

Table 3: *Teacher Rating Correlations*

Method	Traits		Teacher Rating								
			Social Skills						Academic Skills		
			Comm.		Coop.		Engage.		Read	Math	
			1	2	3	4	5	6			
Teach. Rating	Soc. Skill	Com.	1. Conv.	1.0	.718*	.758*	.850*	.247	.463*	.411	.385
			2. Nonverb.		1.00	.787*	.738*	.453	.681*	.649*	.673*
		Coop.	3. Class Part.			1.00	.926*	.130	.396	.652*	.596*
			4. Follow Expect.				1.00	.045	.280	.560*	.567*
		Eng.	5. Group Part.					1.00	.736*	.123	.150
			6. Interact.						1.00	.523*	.502*
	Ac. Skill	Read								1.00	.900*
		Math									1.00

Table 4: *Teacher Rating and Behavioral Observation Correlations*

Methods	Traits			Behavioral Observation							
				Social Skills						Academic Skills	
				Comm.		Coop.		Engage.			
				1	2	3	4	5	6	Read	Math
Teach. Rating	Soc. Skills	Comm.	1. Conv.	.261	.330	.608*	.616*	.441	.431	.029	.123
			2. Nonverb		.276	.442	.477*	.204	.365	.141	.486*
		Coop.	3. Class Part.			.713*	.738*	.352	.360	.166	.379
			4. Follow Expect.				.781*	.408	.363	.195	.247
		Engage	5. Group Part.					.170	.212	-.261	.156
			6. Interact.						.358	-.072	.415
	Ac. Skill	Read								.547*	.649*
		Math									.588*

Table 5: *Teacher and Student Rating Correlations*

Methods	Traits			Student Rating							
				Social Skills						Academic Skills	
				Comm.		Coop.		Engage.			
				1	2	3	4	5	6	Read	Math
Teacher Rating	Social Skills	Comm.	1. Conv.	.065	.056	-.010	.353	.053	-.08	.054	.038
			2. Nonverbal		.413	.233	.314	.281	.053	.172	.269
		Coop.	3. Class Part.			.460*	.684*	.136	-.16	-.064	.282
			4. Follow Expect.				.681*	.091	-.10	-.028	.183
		Engage.	5. Group Part.					.231	.194	.403	.196
			6. Interact.						.182	.321	.204
	Ac. Skills	Read								.185	.082
		Math									.076

Table 6: Behavioral Observation Correlations

Method	Traits		Behavioral Observation								
			Social Skills						Academic Skills		
			Comm.		Coop.		Engage.		Read	Math	
			1	2	3	4	5	6			
Beh. Obs.	Soc. Skill	Comm.	1. Conv.	1.0	.812*	.18	.301	.255	.240	.150	.366
			2. Nonverb.		1.00	.43	.506	.532*	.458*	.020	.214
		Coop.	3. Class Part.			1.00	.883	.700*	.545*	-.25	.043
			4. Follow Expect.				1.00	.716*	.638*	-.01	.004
		Eng.	5. Group Part.					1.00	.818*	-.18	-.07
			6. Interact.						1.00	-.11	-.07
	Ac. Skill	Read								1.00	.44*
		Math									1.00

Table 7: Behavioral Observation and Student Rating Correlations

Method	Traits		Student Rating								
			Social Skills						Academic Skills		
			Comm.		Coop.		Engage.		Read	Math	
			1	2	3	4	5	6			
Beh. Obs.	Soc. Skill	Comm.	1. Conv.	.147	.362	.401	.342	.455*	.576*	.401	.281
			2. Nonverb.		.270	.376	.330	.476*	.467*	.296	.316
		Coop.	3. Class Part.			.270	.468*	.041	-.185	-.204	.380
			4. Follow Expect.				.562*	.166	-.015	-.119	.239
		Engage	5. Group Part.					.023	-.105	-.029	.161
			6. Interact.						.028	-.040	-.078
	Ac. Skill	Read								.119	-.344
		Math									.466*

Table 8: *Student Rating Correlations*

Method	Traits			Student Rating							
				Social Skills						Academic Skills	
				Comm.		Coop.		Engage.		Read	Math
				1	2	3	4	5	6		
Student Rating	Soc. Skill	Comm	1. Conv.	1.00	.746*	.611*	.335	.316	.355	.138	.022
			2. Nonverbal		1.00	.786*	.300	.667*	.558*	.384	.324
		Coop.	3. Class Part.			1.00	.700*	.476*	.336	.118	.506*
			4. Follow Expect.				1.00	.050	-.103	-.200	.424
		Eng.	5. Group Part.					1.00	.807*	.603**	.373
			6. Interact.						1.00	.689**	.232
	Ac. Skill	Read								1.00	.218
		Math									1.00

Table 9 illustrates a condensed depiction of the multitrait-multimethod matrix. Each of the previous tables is represented along with lettered sections designating convergent and discriminant validity values. In Table 9, for example, sections a₁, a₃, and a₅ are analogous to reliability values for the social-skills domains. Sections b₁, b₃, and b₅ represent convergent validity values for social skills. Sections a₂, a₄, and a₆ are analogous to reliability values for the academic-skills domains. Sections b₂, b₄, and b₆ represent convergent validity values for academic skills. Homomethod discriminant validity values are represented in sections c₁, c₂, and c₃, and heteromethod discriminant validity values are represented in sections d₁, d₂, and d₃.

Table 9. Condensed MTMM Matrix

Methods	Teach. Rating		Beh. Obs.		Student Rating		
	Traits	Social	Academic	Social	Academic	Social	Academic
Teach. Rating	Social	a ₁	c ₁	b ₁	d ₁	b ₅	d ₃
	Academic	Table 3	a ₂	Table 4	b ₂	Table 5	b ₆
Beh. Obs	Social			a ₃	c ₂	b ₃	d ₂
	Academic			Table 6	a ₄	Table 7	b ₄
Student Rating	Social					a ₅	c ₃
	Academic					Table 8	a ₆

In order to condense information from each table for ease of interpretation, a measure of central tendency was calculated for each lettered section. The median was used for central tendency as the mean can be influenced by skewed distributions and some level of skewness was anticipated due to low sample size. In some instances, Fisher's z' (Fisher, 1924) was used to assist in calculating the median. Fisher's z' is used to transform Pearson r 's to a normally distributed statistic, with which mathematical calculations can be computed. Following the calculation of the median, the Fisher's z' statistic was transformed back to a Pearson's r for consistency of interpretation.

Social-skills domain. The median correlational values across social-skills domains as outlined previously, were .681 for teacher ratings, .532 for behavioral observations, and .476 for student ratings. In order to assess the consistency of ratings, Cronbach's α (Cronbach, 1951) was calculated for each set of social-skills reliability correlations. Values greater than .9 are considered excellent, values between .7 and .9 are good, values between .6 and .7 are acceptable, values between

.5 and .6 are poor, and values below .5 indicate unacceptable consistency. A value of .878 was obtained for the consistency of teacher ratings, a value of .874 was obtained for behavioral observations, and a value of .853 was obtained for student ratings. All raters/observers displayed a high level of consistency in their ratings.

Convergent validity for social skills. Convergent validity values were obtained using different methods but the same trait (i.e., social skills). The median convergent values were .365 for teacher rating and behavioral observation, .330 for behavioral observation and student rating, and .182 for teacher and student rating. These values provide some evidence of convergent validity, though it is not strong evidence.

Academic-skills domain. The median correlations across academic-skills domains were .90 for teacher ratings, .448 for behavioral observation, and .218 for student ratings. Teachers had, by far, the highest level of reliability for rating academic skills. Cronbach's α was used to calculate consistency for each set of academic skills correlations. A value of .948 was obtained for the consistency of teacher ratings, a value of .619 was obtained for behavioral observations, and a value of .358 was obtained for student ratings. Teachers were very consistent in their ratings, observers were moderately consistent, and students were not very consistent.

Homomethod discriminant validity. Homomethod discriminant values were obtained using the same method but different traits (i.e., social and academic skills). Median discriminant values were .542 for teacher ratings, -.093 for behavioral observation, and .190 for student ratings. Strong discriminant validity was obtained

through behavioral observation and student ratings but not from teacher ratings. Low correlations are desirable in order to indicate that one trait, or construct, does not overlap with another. Moderate correlations were obtained between teacher ratings of social and academic skills, which is not surprising given that some degree of correlation was expected between the two traits.

Heteromethod discriminant validity. Heteromethod discriminant values were obtained using differing methods and different traits. Median discriminant values were .181 for teacher ratings and behavioral observations, .019 for behavioral observations and student rating, and .190 for teacher and student ratings. Strong evidence for discriminant validity was obtained as the values indicate virtually no relationship between academic and social skill ratings across methods.

Test-Retest Reliability

In order to establish reliability between the first and second administrations of rating scales, test-retest reliability was calculated using Pearson product-moment correlations of time 1 and time 2 (T1 and T2, respectively) administrations (administered 2-3 weeks apart) of both teacher and student rating scales. Table 10 displays the obtained correlational coefficients for teacher and student rating scale test-retest reliability.

Table 10: Teacher Test-Retest Reliability

Skill	Reliability Coefficient	
	Teacher	Reliability Coefficient Student
Conversation	.788**	.581**
Nonverbal Communication	.745**	.685**
Classroom Participation	.559*	.731**
Follow Expectations	.813**	.673**

Skill	Coefficient Teacher	Coefficient Student
Group Participation	.577**	.811**
Interaction	.612**	.579**
Reading	.525*	.720**
Math	.737*	.767**

* significant at .05 alpha level **significant at .01 alpha level

All correlations were significant at the .05 level, indicating that any differences between T1 and T2 ratings were likely not due to chance. Many correlational coefficients, however, did not surpass .70, which is used here as an acceptable level of test-retest reliability. Teachers were not adequately reliable at rating students' Classroom Participation, Group Participation, Interaction, and Reading skills. Students were not adequately reliable at rating their Conversation, Nonverbal Communication, Follow Expectations, and Interaction skills.

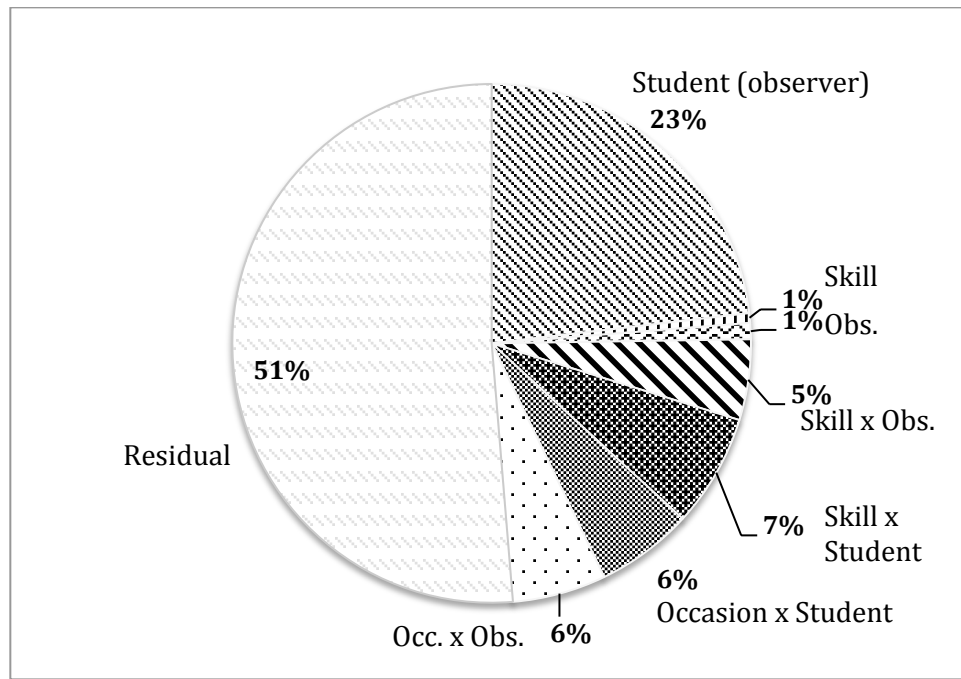
Generalizability and Decision Studies

The VARCOMPS procedure was used to compute the variance components analyzed in the G2.sps SPSS program developed by Mushquash and O'Connor (2006, revised 2012). The Matrix-End Matrix procedure was used to read the variance components according to the specifications of the design, and G-theory results were obtained. Results of this G-study are presented in Table 11, which lists the sources of variation, the variance components, and the proportions of total variance explained by each facet; Figure 1 presents the proportions of variance explained by each of these sources in a circle graph.

Table 11: *G-Study Results*

Source of Variation	Variance Component	Proportion of Variance
Student (Observer)	134.239	.234
Skill	3.175	.006
Occasion	0.000	.000
Observer	6.730	.012
Skill x Occasion	0.000	.000
Skill x Observer	30.630	.053
Occasion x Observer	32.990	.058
Skill x Student (Observer)	39.736	.069
Occasion x Student (Observer)	36.034	.063
Residual	289.161	.505
Total	--	1.000

Figure 1: *Percents of Variance Explained*



The overall, relative G-Coefficient, which describes the universal reliability of the measure, was .936. This coefficient indicates that approximately 94% of the variance was accounted for with approximately 6% of the variance representing error. The residual term accounted for the greatest portion of variance (i.e., 51%). Students (i.e., the object of measurement, accounting for nesting within observer) accounted for 23% of the variance. The interactions of Occasion-by-Observer, Occasion-by-Student, Skill-by-Student, and Skill-by-Observer accounted for smaller proportions of 6%, 6%, 7%, and 5%, respectively.

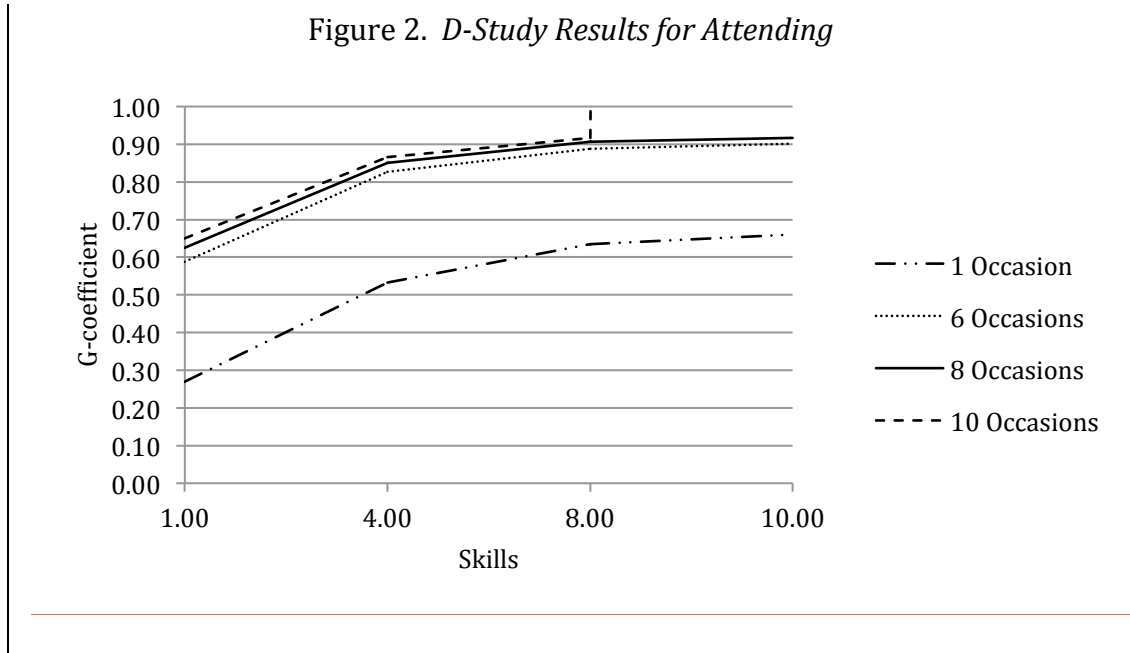
Next, D-studies were conducted in order to estimate how varying levels might affect the reliability of each facet. These D-study results are presented in Table 12.

Table 12. *D-Study Results*

Observer	Skill	Occasion			
		1	6	8	10
1	1	.269	.588	.625	.650
	4	.532	.827	.851	.866
	8	.635	.888	.906	.917
	10	.661	.901	.917	.928
4	1	.595	.851	.870	.881
	4	.820	.950	.958	.963
	8	.874	.969	.975	.978
	10	.886	.973	.978	.981
5	1	.648	.877	.893	.903
	4	.850	.960	.966	.970
	8	.897	.975	.980	.982
	10	.907	.978	.982	.985
7	1	.720	.909	.921	.929
	4	.888	.971	.976	.978
	8	.924	.982	.985	.987
	10	.932	.984	.987	.989

Figure 2 provides graphic illustration of the relative G-coefficients for level-1 Observer relative G-coefficients. Closer examination is given to the level-1 Observer D-study results because it most closely resembles scenarios applicable to a practicing school psychologist.

Figure 2. *D-Study Results for 1 Observer*



Interobserver Agreement

Interobserver agreement was calculated for observer pairs consisting of the primary researcher and each undergraduate assistant. A secondary observer was randomly assigned to 19 out of 20 student participants. As stated previously, the fourth occasion was used for the interobserver recordings as the secondary observer simultaneously observed each skill with the student's primary observer. In other words, 19 students were observed on each of the skills by two observers on the fourth occasion of observation. Interobserver agreement was unable to be

calculated for one student who was chronically absent at the end of the data collection period, and a fourth occasion/interobserver observation was not obtained. Interobserver agreement was calculated using SPSS Crosstabs function, which produces a Kappa statistic for level of agreement. According to Cohen (1960), Kappa values lie between -1.00 and 1.00, with 0 indicating chance agreement, positive values indicating greater than chance agreement, and negative values indicating less than chance agreement. Landis and Koch (1977) categorized Kappa values from 0.41 to 0.60 as moderate and values above .60 as substantial levels of agreement. Table 13 displays the level of agreement for each skill pair.

Table 13. *Interobserver Agreement Kappa Values*

Observer Pair	Kappa Value
Observer 1 and Observer 2	.157
Observer 1 and Observer 3	.032

Primary and secondary observers displayed agreement around chance levels. Results previously described in the G study indicated that rating style of the observers had minimal influence on scores, but that the observer interaction with other facets accounted for approximately 11% of the measurement variance when combined.

Descriptive Analyses

A series of five, multivariate analysis of variance (MANOVA) tests were conducted in order to assess any score differences in the behavioral observation data based on demographic categories. Dependent variables included scores on

each of the six social skills that were observed. Grade included seven groups from grades 6 through 12, Gender included two groups (male and female), Ethnicity included three groups (White, African American, and Hispanic), SES included two groups (low and medium to above, as previously described), and Diagnosis included five groups (Mood Disorder, Autism Spectrum Disorder, Attention Deficit Hyperactivity Disorder, Anxiety Disorder, and Post-Traumatic Stress Disorder). As each student met more than one diagnostic category, the primary diagnosis was used as the descriptive factor. Table 14 displays the results of MANOVA analyses.

Table 14. *Results of MANOVAs for Demographic Factors*

Factor	Wilks' λ	F	df	Error df	p	η^2
Grade	.023	1.432	36	37.891	.139	.467
Gender	.870	.324	6	13.000	.913	.130
Ethnicity	.627	.525	12	24.000	.877	.208
SES	.596	1.470	6	13.000	.263	.404
Diagnosis	.134	1.059	24	32.607	.433	.395

There were no significant MANOVAs for the demographic factors, indicating that participants did not differ significantly in their performance of social skills based on the given descriptive factors. It should be noted that group size within each demographic factor was not equally represented, with some groups often much larger than others. For example, the gender of the sample population was largely male ($n = 15$). It is possible that differences in skill performance may have been seen if the number of participants had been larger and the distribution across groups within factors more similar.

Chapter IV: Discussion

Students' social functioning is important to both the social and academic success of children and youth (Cappadocia & Weiss, 2011; Cook et al., 2008; Gresham et al., 2004). Indeed, outcomes for students with poor social skills often include behavioral, emotional, and academic difficulties (Cook et al., 2008). Often, the only access that children and adolescents have to mental-health services, including social-skills instruction, is in a school-based setting (Hoagwood & Johnson, 2003). Unfortunately, social-skills interventions in schools are often ineffective (Cappadocia & Weiss, 2011; Gresham et al., 2001), which may be related in some ways to measurement and assessment issues (Matson & Wilkins, 2009). There is no single, agreed-upon definition of social skills, and measurement strategies are numerous and varied (Crowe et al., 2011; Matson & Wilkins, 2009). The present study evaluated a set of skills with the purpose of identifying a homogenous skillset, which could be used for social-skills intervention, and a measurement strategy that could be used for progress monitoring of skills and decision-making regarding social-skills instruction.

This study used a multitrait-multimethod approach to assess a social-skills construct, or identify a homogenous group of skills, and generalizability theory to evaluate a behavioral-observation tool. The multitrait-multimethod approach used academic skills as a discriminant trait for comparison with social skills. Behavioral observation, teacher ratings, and student ratings were used for methods of obtaining participant data on both traits. Convergent and discriminant validity were assessed through correlations among all traits and methods.

G theory was chosen for use in this study because of the benefits it has over the traditional approach of CTT. The present study used G theory to evaluate multiple sources of variance, or facets (e.g., observer, occasion, skill), and obtain an overall value of reliability that accounts for those sources of variance. G theory was also used to predict the reliability of a behavioral observation tool given alternative levels of each facet, different from those used in the original study.

Psychometric Findings

The current study used a multitrait-multimethod approach to assess the validity of a social skills construct, using social skills and academic skills as traits and behavioral observation and teacher and student rating scales as methods. Additionally, this study used G theory to examine the reliability of an observational tool to observe student performance of social skills. The measurement strategy included student nested within observer as the object of measurement and skill, occasion, and observer as facets. The present study used a nested design, meaning that each facet does not occur at each level with every other facet. Some facets may occur only at some levels and not at others. For example, one might have a study where some skills are observed on particular occasions and others are not. In the current study, students were nested within observers; observers were randomly assigned only to particular students and did observe every student. A completely crossed design would require each observer to observe every student on all skills for each occasion. A discussion of each study and the attributed follow-up analyses follows.

Multitrait-multimethod Study

Pearson product moment correlations were calculated for each trait-method combination to assess convergent and divergent validity of the social-skills construct. Results yielded six trait-method combinations, through which values were obtained for the reliability of social and academic skills, convergent validity, and discriminant validity of social skills. Median values were calculated in order to provide a summary value for each trait-method matrix. Cronbach's α was also calculated for social and academic skills reliability matrices to assess the consistency with which skills were rated.

Social-skills domain. The obtained median values were .681 for teacher ratings, .532 for behavioral observations, and .476 for student ratings. These values were somewhat unexpected, as behavioral observation was anticipated to have the lowest reliability. Although median values are moderate, they are lower than expected. Rater inconsistency cannot be used to explain these results, as the consistency of ratings was found to be above .85 for all methods, indicating that raters were consistent across methods. It is likely that reliability for teacher and student ratings would be higher given a larger number of items to assess each skill. Social skills consisted of 3 to 4 items on each rating scale, which may not be sufficient to obtain a high degree of reliability for each skill. Additionally, the cohesiveness of skills may also have had an impact on reliability. As anticipated, some skills appeared to be more highly correlated than others. Class Participation and Follow Directions (Cooperation) appeared to be highly correlated across methods, obtaining a value of .926 on teacher ratings, .883 on behavioral

observations, and .700 on student ratings. The correlation between skills varies across methods, making the median value appropriate for interpretation, but the close relationship of the Cooperation skills is consistent.

Convergent validity for social skills. Values for convergent validity were obtained through correlations of social skill ratings using different methods. Median values were .365 for teacher rating and behavioral observation, .330 for behavioral observation and student rating, and .182 for teacher and student rating. Median correlations were small in magnitude, but again, moderate to high correlations were obtained for the Cooperation skills of Classroom Participation and Follow Expectations. Additionally, Group Participation and Interaction (Engagement) were only loosely correlated across methods.

Academic-skills domain. Median values were .90 for teacher ratings, .448 for behavioral observations, and .218 for student ratings. Here, consistency of ratings may have influenced reliability values. Again, Cronbach's α was used to assess consistency among ratings. A value of .948 was obtained for teacher ratings, .619 for behavioral observations, and .358 for student ratings. Teacher ratings had a high level of consistency and reliability, behavioral observations had a poor consistency and moderate reliability, and student ratings had unacceptable consistency and low reliability. Inconsistent ratings for academic skills, however, may not be unexpected. Reading and math skills were grouped together as a single construct, but students who perform well on one skill may not necessarily perform well on the other. These results indicate that these skills may not necessarily be grouped together reliably. Teachers appear to rate students as performing

similarly, or consistently, on both behaviors, whereas behavioral observations and student ratings assess skill performance independently of one another.

Homomethod discriminant validity. Discriminant values were obtained for same method (e.g., teacher ratings) and differing traits (i.e., social and academic skills). Median values were .542 for teacher ratings, -.093 for behavioral observations, and .190 for student ratings. Low discriminant values, such as those obtained for behavioral observations and student ratings provide evidence for the divergence of constructs. In other words, social and academic skills are separate and distinct constructs. Teacher ratings were moderate, however. One reason for this may be that teachers rate students as performing similarly across skills. For example, a student who shows good performance in academic areas may be rated equally well on social skills, and those who perform poorly in academic areas may be rated low on social-skill performance. Teachers may see social skills in their relation to classroom performance, whereas students may not make that connection and behavioral observations may not reflect such a relationship.

Heteromethod discriminant validity. Discriminant values were obtained for different methods (e.g., teacher ratings and behavioral observations) and different traits (i.e., social and academic skills). Median values were .181 for teacher ratings and behavioral observations, .019 for behavioral observations and student ratings, and .190 for teacher and student ratings. As anticipated, these values are the lowest of the trait-method combinations. These values provide strong evidence for the divergence of social and academic skills.

Thus, overall, as shown in Table 15, the square root of homotrait-homomethod correlations (i.e., “*a*”s) for the social skills domains generally exceeded convergent correlations (i.e., “*b*”s), which, in turn, exceeded discriminant homomethod correlations (i.e., “*c*”s), which, in turn, exceeded discriminant heteromethod correlations (i.e., “*d*”s). Of course, this is the pattern that would be expected according to Campbell and Fiske’s paradigm to establish construct validity. An important exception is that teacher ratings of social and academic skills were unusually high (i.e., .542 versus -.093 for behavioral observation and .190 for student methods of assessment). It may be that teachers in this specialized school setting link social and academic skills more closely than would be seen in a typical school setting.

Table 15. *Median Multitrait-Multimethod Correlation Coefficients for Social Skills*

		Social Skills			Academic Skills		
		Teacher	Behavior Obs.	Student	Teacher	Behavior Obs.	Student
Soc. Skill	Teacher	<i>a</i> = .681	<i>b</i> = .365	<i>b</i> = .182	<i>c</i> = .542	<i>d</i> = .181	<i>d</i> = .190
	Beh. Obs.		<i>a</i> = .532	<i>b</i> = .330		<i>c</i> = .093	<i>d</i> = .019
	Student			<i>a</i> = .476			<i>c</i> = .190

Test-retest reliability. Rating scales were administered on two occasions (2 to 3 weeks apart) to teachers and students. Reliability coefficients ranged from .525 to .813 for teacher ratings and from .579 to .811 for student ratings. Correlations for each skill were significant from time 1 to time 2. Many of the reliability estimates, however, did not reach a level of .70, which would indicate a good level of reliability for a measure. The SSIS Rating Scales Manual (pp. 67-68) identifies reliability coefficients ranging from .68 to .86 across skills with a mean adjusted

coefficient of .82 for the teacher form. For the student form, values ranging from .59 to .81 across skills with a mean adjusted value of .71. Reliability coefficients may be lower in the present study due to the number of skills assessed. The SSIS manual provides reliability values for composite skills, whereas the present study assessed the skills comprised within. For example, Conversation and Nonverbal Communication were assessed as separate skills (3-4 items each) rather than the composite skill of Communication (6-8 items). Reliability estimates are likely to be lower due to a lower number of items representing each skill.

Generalizability Study

As previously discussed, the G study provides a coefficient that describes the global reliability of a measure, variance component values and percentages of variance accounted for by each facet, and residual error. The D study provides variance component values and associated G-coefficients for alternative measurement strategies given varying levels of each facet. A discussion of the G study and D study results from the current study follow.

Relative G-coefficients were reported as a measure of overall reliability for the measure, as well as for the D study, because they are most applicable for the applied use of the behavioral measure of interest. Generalizability theory can be used for the purposes of making relative decisions and absolute decisions. Both G and D studies provide relative and absolute coefficients applicable for each decision-making purpose, respectively. Relative decisions are those concerning an individual's performance compared to others, whereas absolute decisions are those concerning an individual's performance compared to a specific criterion regardless

of others' performance. The behavioral observation tool used in this study could be used to make both types of decisions. For example, relative decisions would be useful in identifying students with similar skill deficits for the purpose of placing them in similar instructional groupings. During a social-skills intervention, absolute decisions would be useful to influence individualized instruction depending on whether or not a student has met criteria for mastering specific goals. The relative G-coefficient is analogous to the reliability coefficient in CTT and is a more accurate indicator of reliability than the absolute Phi-coefficient of dependability (Shavelson & Webb, pp. 93). Thus, relative G-coefficients were reported for the purpose of relative interpretations in this study.

The overall, relative G-coefficient (used for decisions based on the relative standing of comparison to others) of the measure was .936, an excellent level of reliability for a behavioral measure. The residual error term accounted for the greatest portion of variance (i.e., 51%), and includes all 3-way interactions between facets that cannot be statistically partialled out. Students nested within observer accounted for the next largest portion (i.e., 23%). The interaction effects accounted for small portions of variance separately, but approximately 24% when combined. Skill and Observer only accounted for approximately 1% of the variance each, so they are not discussed further. The interaction effect of skill and observer would be of interest in identifying the effect of individual student performance for particular skills. A low percentage of variance accounted for, as is seen here, indicates that the measure may assess effects for the interaction of student and individual skill with low reliability.

Utilizing G theory provided a better assessment of measurement reliability than would be available if using CTT, as the percent of variance accounted for by each facet would have been otherwise attributed to random error. Ideally, the student facet would contribute the largest portion of variance, but here, it accounts for approximately a quarter of the variance. The student facet is of interest because it pinpoints the amount of the effect that can be attributed to a particular student (i.e., what portion of variance is attributed to the student, themselves, and not to outside factors). The student facet, combined with all 2-way interaction effects accounts for approximately half of the variance, the other half being accounted for by residual error and all 3-way interactions among facets. Here, the proportion of variance accounted for by interaction effects is compelling. These results speak to the true nature of behavioral data, which can vary markedly across occasions and environments. Also, given the clinical nature of the sample population, social-skills performance may have been even more variable than a typical student population.

D-study. The D study results showed reliability levels for the behavioral measure given alternative levels of each facet. Discussion of results for alternative levels of occasion and skills for a single observer follows. A single observer is considered here, as it most closely resembles the scenario of a typical practicing school psychologist. G-coefficients were calculated for 1, 4, 8, and 10 skills in combination with 1, 6, 8, and 10 occasions. With a single observer, one skill and one occasion had insufficient reliability with a value of .269. A combination of 4 skills and 6 occasions yielded a good level of reliability of .827. Combinations of 8 skills and occasions, as well as 10 skills and occasions yielded reliability coefficients above

.90. The reliability increased greatly between 1 and 6 occasions, but appeared to taper off quickly after the combination of 4 skills and 6 occasions. Optimal reliability was found with a combination of 10 skills and 10 occasions (i.e., .928). However, it may not be worth the cost in time and resources, when an adequate level of reliability could be obtained using fewer occasions and skills.

Interobserver agreement. Cohen's Kappa was used to calculate interobserver agreement. As stated previously, Kappa values range between -1.00 and 1.00 with 0 indicating chance agreement. Positive values indicate greater than chance agreement and negative values indicate less than chance agreement. Kappa values were found to be $K = .157$ for the first observer pair and $K = .032$ for the second. Agreement for both pairs, or among the three observers, was close to chance levels. These findings may have resulted from a combination of the way that data were represented and the way that Kappa is calculated. Generally speaking, Kappa uses the frequency of agreement between observers in designated categories (e.g., number of yes/no observations per rater). This type of calculation may underrepresent the level of agreement between observers, as the range of scores in the present study is much larger than would be accounted for by a yes/no type of response. For example, every participant was rated from 0-100% on each skill, so they could obtain values anywhere from 0-100. If observer 1 rated student A as completing a skill with 80% accuracy and observer 2 rated student B as completing the same skill with 85% accuracy, Kappa calculations may not account for this as a "match" between observers even though both values fall within a similar range of scores. Given this consideration, and the G study results that found the observer

facet to be of little influence on variance, the Kappa values are viewed to be an underrepresentation of interobserver agreement.

Cultural Considerations

Information on multiple demographic factors was collected for each participant, including, age, grade, gender, SES, ethnicity, and primary mental-health diagnosis. Individual analyses were conducted for each demographic factor to determine if scores may have varied among factors. No significant effects were found for any demographic factor. However, sample size was small and the size of groups within factors was often unbalanced. Given the large effect sizes found for grade ($\eta^2 = .467$), SES ($\eta^2 = .404$), and diagnosis ($\eta^2 = .395$), it is possible that a larger sample size and more balanced level of participants across within-factor groups could reveal differences. Additionally, it should be noted that the majority of participants exhibited significant social-emotional difficulties as a result of mental-health disabilities. Further examination of potential differences in skill performance based on diagnosis would be beneficial to the field in application to the building of more effective social-skills interventions.

Implications

This study revealed a cohesive set of skills that may be identified as “social skills” and reasonably measured together as a homogenous group, or skillset. Conditions varied across methods, and some skills were more closely correlated than others, but results largely supported the assessed skills as a cohesive and distinct skillset. As some skills were more closely linked than others (i.e., Classroom Participation and Follow Expectations were more highly correlated than

Conversation and Follow Expectations), particular skills may be viewed as heterogenous. Taken together, however, they represent a skillset, or homogenous group, to be termed here as a social skills construct. As no single, unified definition of social skills has been established, a method for selecting homogenous skillsets would be beneficial. In this study, skills were selected from subscales of the SSIS. A wide variety of validated instruments, similar to the SSIS, are comprised of multiple sets of skills. It is probable that alternative social skillsets could be identified, using a number of validated instruments, and targeted for intervention. Various skills are often grouped together idiosyncratically and termed “social skills.” This study showed how a method can be employed to identify similar skills so that a more targeted intervention may be developed, and hopefully, be more effective for students.

The present study demonstrated the usefulness of employing G theory when developing a behavioral measure. Multiple behavioral-observation measures using technology such as iPhone applications exist, but few, if any, have conducted similar studies in order to validate said measures. This study presented an illustration of how G theory can be used assess the adequacy of a specified measurement strategy as well as the adequacy of alternative strategies with varying levels of facets. Similar studies could be conducted on other behavioral observation measures in order to assess their adequacy for the same purpose.

The behavioral observation tool, and similar tools once validated, could be used in multiple stages of the intervention process. First, it could be used as a screening measure to identify students with similar skill deficits (relative

comparison). Second, after students with similar skill deficits have been grouped together for instruction, it could be used as an observational tool to collect progress-monitoring data of student performance of social skill. Third, information collected during progress monitoring could be used to influence the differentiation of instruction for individuals (absolute comparison), such as by gauging the completion of criteria or benchmarks that indicate the mastery of particular skills. These uses link directly to the implementation of an RTI format for social skills. As discussed previously, the RTI process is frequently used in schools for academic instruction, but is rarely seen for use in social-emotional instruction. This research adds to this area of study by establishing a reliable and feasible measurement strategy to assist in all stages of the RTI process for social-emotional education. Salvia and Ysseldyke (2004) suggested that reliability coefficients of .90 or higher are recommended for instructional decision-making purposes and coefficients of .70 or higher are recommended for screening purposes. The obtained reliability for the measure used in this study was above .90 and could, therefore, be used for both screening and instructional decision-making purposes.

As other behavioral measurement tools are developed for a similar purpose (Cummings et al., 2008; Stichter et al., 2012), consideration should be given to the implications of using behavioral observation as a primary data collection strategy. Hintze and Matthews (2004) suggested momentary time sampling to be a more favorable observation strategy than partial or whole-interval methods as it results in smaller estimation errors. The authors also cautioned the overuse of behavioral observation, as it is less reliable than commonly believed. As shown in the present

study, a number of observations may need to be conducted on each skill in order to obtain an adequate level of reliability. School psychologists may find the amount of time required to observe skills reliably to be unmanageable in addition to other job responsibilities. Baer, Harrison, Fradenburg, Petersen, and Milla (2005) stressed the importance of operational definitions of target behaviors and time and setting of observations when assessing behavioral data. As illustrated in the present study, variables such as skill and occasion may interact and produce an impact on the behavioral outcome. Operational definitions are important to the consistency of skill ratings and should be established clearly before any observations are conducted.

Limitations

First, although the multitrait-multimethod approach used in this study provided useful information for the establishment of a social skills construct, it provided so much information that findings needed to be condensed for ease of interpretation. When information is condensed in such a way, some robustness of the overall picture may be lost.

Second, the present study used G theory to assess the usefulness of a measurement strategy because it is less restrictive than CTT and considered multiple facets of the measurement design. Although multiple facets and 2-way interactions between them were assessed, there are still variables left accounted for. Variables such as setting, time of day, and activity were not controlled for or evaluated in the present design and may have played some role in the outcome of student performance on specific skills, as suggested by the G-study interaction

effects. Currently, the procedure for assessing more than three facets in a generalizability design is unavailable, and holding environmental variables constant, or controlling the facets used in this study another way, may be the only options for assessing their potential impact on skill performance. Holding environmental variables constant may also reduce the amount of variance attributable to random error.

Third, a limitation in the observational method, primarily the interval length used in the observations, should be noted. Because beginner observers were used, a longer interval was selected (i.e., 30 seconds) in order to obtain a more accurate score. A longer interval was used to reduce the effort needed to track interval length and in hopes of obtaining accurate scores for the appropriate interval. More experienced observers would be able to use 10 or 15-second intervals while keeping track of time and student performance, which can provide a closer approximation to the percentage of time spent engaging in the specified skill. Additionally, 15-minute observations were conducted in order to maximize the number of students who could be observed during the limited time frame for data collection, which may not be an adequate amount of time to obtain a representative sampling of student behavior on some skills.

Finally, due to the nature of the school setting where the study took place, sample size was limited. Although the number of data points collected for each participant made the analyses of interest for the present study appropriate, skill performance differences based on multiple cultural factors could not adequately be assessed as a result of the small sample size. Additionally, the sample size was

unbalanced, where within-factor groups often had only a few members. The sample also represented a clinically based, low SES, primarily male student population, which may limit the generalizability of research findings, somewhat, to typical student populations. It is possible that the participant population is qualitatively distinct from the student populations of other school settings.

Future Directions

The present study provided a number of results from which future research and practice can be based. A multitrait-multimethod approach was used to identify a cohesive set of social skills, which could be targeted for social-skills intervention. Although this approach provided a variety of useful results, it provided more information than could reasonably be interpreted and was condensed for ease of interpretation. Attempts to identify other cohesive skillsets could utilize alternative approaches, such as structural equation modeling, which may better organize findings and provide a structure to the results to ease interpretation.

By using G theory, a reliable measurement strategy was developed that could be used in schools. However, limitations exist for the analysis of more than three facets in their contribution to measurement variance. Future research could be conducted to account for the impact that environmental factors may have on skill performance. One method may be to conduct a G-study using a single observer and skill (e.g., control for observer and skill), and use environmental factors as facets. Researchers may also consider examining the potential impact of teacher experience on student behavior. For example, one could examine whether more experienced teachers report fewer behavior problems in their classroom, as they

may be more comfortable with students and have more effective behavioral strategies.

It is hoped that school practitioners may utilize the findings of this study to inform their practice by identifying cohesive groups of skills to target for intervention and using behavioral observation measures in a reliable way as part of a three-tiered instructional model. An RTI-based approach to social-skills instruction might include screening students in order to place students with similar skill deficits into groups for instruction, progress monitoring of skills using semi-weekly behavioral observations of students in multiple environments, and differentiating instruction based on observed student progress. In other words, future research could develop a method, or “best practice” for implementing a behavioral-observation tool in such a manner. Additionally, future research could be used in an experimental way to determine if the use of behavioral observation measures in a three-tiered format has an impact on intervention effectiveness. Future research should also investigate the use of behavioral-observation measures, such as the one used in this study, in multiple school settings with differing student populations in order to assess their appropriateness for various populations.

Summary and Conclusions

Social skills play an important role in student success in schools, both socially and academically (Cappadocia & Weiss, 2011; Cook et al., 2008; Gresham et al., 2004). Historically, social-skills interventions have lacked effectiveness in the way of generalization of skills to settings outside of the instructional environment (Gresham, 2010). Ineffectiveness of interventions often may be related to

assessment and measurement issues (Matson & Wilkins, 2009). In schools, attention has started to be given to the development of progress-monitoring tools for social-emotional interventions (Cummings et al., 2008; Stichter et al., 2012); these tools have attempted to utilize behavioral observation in a reliable manner. This study used multitrait-multimethod and generalizability approaches to develop such a measurement strategy.

The present study demonstrated an approach to identifying skillsets for targeted intervention by evaluating the convergent and divergent validity for social and academic skills. A structured approach to identifying cohesive target skills strengthens the underpinnings of social-skills interventions by providing an evidence-based approach rather than relying on a heterogeneous grouping that may interfere with measurement reliability.

The current study also demonstrated the usefulness of G theory for developing a multifaceted measurement strategy for behavioral observation. G theory expands the CTT perspective by including multiple facets to account for aspects of variance in addition to random error. In addition, G theory can be used to assess how different levels of each facet might affect the measure's reliability in alternative measurement scenarios. In this study, the skills of interest were Conversation, Nonverbal Communication, Classroom Participation, Follow Directions, Group Participation, and Interaction. The measurement design used students nested within observers as the object of measurement and skill, occasion, and observer as facets.

Results indicated that a cohesive set of skills had been identified, but that the findings may have been stronger if more items had been used on rating scales to assess each skill (providing a more reliable estimate of each skill). G-study results indicated a good level of reliability for the measure, and that approximately half of the error could be attributed to students and the 2-way facet interactions combined. D-study results indicated that an adequate level of reliability could be obtained using multiple observers and a moderate number of occasions, but that the number of occasions would need to be increased for a single observer to obtain adequate reliability.

This study established a method for assessing a cohesive skillset for intervention and established a reliable measurement strategy that lends itself to multiple decision-making purposes in the intervention process. It is hoped that social skills instructors could utilize these methods to create better-planned interventions and use progress-monitoring practices. By creating groups with homogenous skillsets and using progress monitoring to inform the decision-making process regarding student social skills performance, the effectiveness of social-skills interventions may be improved and the social-emotional functioning of students may be improved. Future research should seek to assess the reliability of behavioral measures, similar to the one used in this study, with different school settings and students from multiple backgrounds. Future research should assess the usefulness of behavioral measures to inform the decision-making process and what impact their use may have on the effectiveness of interventions.

Appendix A: Expanded Multitrait-Multimethod Matrix

Methods	Traits	Teacher Rating						Behavioral Observation						Student Rating							
		Social Skills			Academic Skills			Social Skills			Academic Skills			Social Skills			Academic Skills				
		Communication	Cooperation	Engagement	Read	Math	Engagement	Communication	Cooperation	Engagement	Read	Math	Engagement	Communication	Cooperation	Engagement	Read	Math	Engagement		
Teacher Rating	Social Skills	1. Conversation	a	b	b	b	c	c	w	x	y	y	z	z	w	x	y	y	y	y	
		2. Nonverbal Communication		b	b	b	c	c	w	x	y	y	z	z	w	x	y	y	y	y	
		3. Classroom Participation			a	b	c	c	w	x	y	y	z	z	w	x	y	y	y	y	
		4. Follow Expectations			b	b	c	c	w	x	y	y	z	z	w	x	y	y	y	y	
		5. Group Participation				a	c	c	w	x	y	y	z	z	w	x	y	y	y	y	
		6. Interaction					c	c	w	x	y	y	z	z	w	x	y	y	y	y	
	Academic Skills	Reading																			
		Math																			
		1. Conversation						a	b	b	b	c	c	w	x	y	y	y	y	y	
		2. Nonverbal Communication						b	b	b	b	c	c	w	x	y	y	y	y	y	
		3. Classroom Participation							a	b	b	c	c	w	x	y	y	y	y	y	
		4. Follow Expectations							b	b	b	c	c	w	x	y	y	y	y	y	
Behavioral Observation	Social Skills	5. Group Participation					a	b	b	b	c	c	w	x	y	y	y	y	y		
		6. Interaction									c	c	w	x	y	y	y	y	y		
		Reading																			
		Math																			
		1. Conversation																			
		2. Nonverbal Communication																			
	Academic Skills	Communication																			
		Cooperation																			
		Engagement																			
		Reading																			
		Math																			
		Engagement																			
Student Rating	Social Skills	Cooperation																			
		Engagement																			
		Reading																			
		Math																			
		Engagement																			
		Reading																			
Academic Skills	Reading																				
	Math																				
	Engagement																				
	Reading																				
	Math																				
	Engagement																				

Appendix B: Parent/Caregiver Consent Form

**The University of Rhode Island
Department of Psychology**



PARENTAL PERMISSION FORM FOR RESEARCH

Your child has been invited to take part in a research project described below. My name is Monica Mabe and I am a graduate student at the University of Rhode Island (URI) and will be conducting this research project with Professor W. Grant Willis, a faculty member at URI. I am asking for permission to include your child in this study because he/she is a student in one of the classrooms selected to participate in this study. This research project will begin in November and be completed in the school by February. This study has been approved by the Executive Board of South Coast Educational Collaborative and administrators of the school.

Description of the project:

Until recently, there have not been many tools available for measuring student behaviors and tracking behavioral progress in areas such as social skills. The purpose of this project is to see if one of the recently developed behavioral measures can be used for observing how adolescents engage in various social skills such as working in groups or following classroom expectations. In other words, the goal is to assess a measurement strategy that can accurately measure and monitor the progress of student social skills.

What will be done:

If you allow your child to participate, here is what will happen: A student from the University of Rhode Island (URI) will be assigned to your child's classroom and observe them during their regular scheduled day. Your child will not be asked to leave the classroom or speak to the URI observer alone. The URI student is only interested in observing different social skills used by your child in the classroom and how they happen during a regular day. The URI students will be observing multiple students in the classroom, so your child will not be identified or singled-out as being observed. Your child will also be asked to complete a form that asks him/her to rate his/her performance of various social skills. Students will be able to complete this form with their counselor or the graduate student researcher.

Risks or discomfort:

There are no risks or discomfort involved for your child in this project. It will be explained to them that there will sometimes be a person from URI observing the

classroom so that they are comfortable and know who will be visiting their classroom.

Benefits of this study:

Although there may be no direct benefit to your child for participating in this project, the school will benefit greatly from the information that will be collected. The information from this project will help personnel at the school improve their data collection procedures so that they are more accurate and meaningful. Results of this study will be made available for viewing online through the ProQuest library search engine. A hard copy of results can also be viewed at the University of Rhode Island library after the completion of the study.

Confidentiality:

Your child's part in this study is confidential. All information will be stored electronically in the online system connected to the behavioral measure, which requires an account with a secure login and password that is only issued to a few individuals at the school. Only individuals directly involved in the study will have access to the secure information. After all of the information is collected, an identification number will be used in alternative to student names; all names will be deleted and there will be no way of tracking any collected information back to an individual student.

Decision to quit at any time:

Students will be given the opportunity to decide whether or not to participate in this project. Their decision to participate will not affect your or their relationship with South Coast Educational Collaborative. Your child will have the right to stop participating at any time. You have the right to withdraw your permission for your child to participate at any time.

Rights and Complaints:

If you are unhappy with the way this study is happening in your child's classroom, you may talk about your complaints with Professor W. Grant Willis (401) 874-4245 or with Graduate Student, Monica Mabe (508) 336-8212 ext. 27, both from URI. A key staff member at the school to contact regarding this project is Program Director Charles Seekell (508) 336-8212. Please feel free to contact any of the individuals listed with further questions you may have about this research project. In addition, if you have questions about your child's rights as a research participant, you may contact the office of the Vice President for Research, 70 Lower College Road, Suite 2, University of Rhode Island, Kingston, Rhode Island, telephone: (401) 874-4328.

You have read this Permission Form. Your questions have been answered. Your signature on this form means that you understand the information and you agree to allow your child to participate in this study. Thanks so much for your attention to this.

I agree to allow my child to participate in this study ___ Yes ___ No

Signature of **Parent**

Signature of Researcher

Typed/printed Name of **Student**

Typed/printed name

Date

Date

Please sign both consent forms, keeping one for yourself

Appendix C: Student Assent Form

**The University of Rhode Island
Department of Psychology**



STUDENT ASSENT FORM

My name is Monica Mabe. I am a graduate student at the University of Rhode Island (URI). I am inviting you to participate in a research study because I am trying to learn more about social skills and students your age. I will explain about the study, but you can ask questions by contacting me later if you want to know more.

Description of the Project:

Recently, a lot of behavior observational tools have become available for use on phones and tablets. Part of this project is to see if one of these observation tools can be used for observing social skills. Social skills are things that people do when they interact with others, like having conversations, working in groups in class, or inviting someone to join a game. The other part of this project is to see if social skills can be measured equally well with paper forms called rating scales and through observation.

What will be done:

If you agree to be in this study, you will be asked to complete two forms about social skills. The forms take about 10 minutes to complete and ask you to rate how you do different things like getting along with others or asking for help when you need it. You will be able to complete this form with a staff member at school in case you have any questions about the items. Sometimes there will be a person from URI in your classroom. They will be looking to see how you and other students act in normal classroom situations. You will not know if they are there to observe you or if they are there to observe other classmates.

Risks or discomfort:

There are no risks or discomfort involved in this project. There will sometimes be a person from URI observing the classroom. Many students in each class will be participating and none of your classmates will know who else in the classroom has agreed to participate in the project.

Benefits of this study:

Although there may be no direct benefit to you for taking part in this study, we may learn more about measuring social skills and students your age. You will have the opportunity to express your opinion about your social skills and learn more about yourself at the same time. Results of this study will be made available for viewing online through the ProQuest library search engine. A hard copy of results can also be viewed at the University of Rhode Island library after the completion of the study.

Confidentiality:

No one else will know if you were in this study and no one else can find out what answers you gave on the social skills form. All of the information from this project will be stored in a locked office on the URI campus.

Decision to quit or not participate at any time:

I will also ask your parent/guardian to give their permission for you participate in this project, but even if your parent/guardian says “yes”, you can still decide not to do this. If you do decide to participate, you can always drop out of the study at any time. No one will be upset if you don’t want to participate or even if you change your mind later and want to stop. If you want to quit the study, just let me know or ask one of your parents/guardians to call me.

Remember, you can ask any questions you may have about this study. If you have a question, you can call me at (508) 336-8212 ext. 27 or Professor W. Grant Willis, who is working with me on this project, at (401) 874-4245. You can also talk to Charles Seekell regarding this project.

Signing your name at the bottom of this form means that you have read or listened to what it says and you understand it. Signing this form also means that you agree to participate in this study and your questions have been answered.

I agree to participate in this study ___ Yes ___ No

Signature of participant

Signature of Researcher

Typed/printed Name

Typed/printed Name

Date

Date

Please sign both assent forms and keep one for yourself

Appendix D: Teacher Consent Form

**The University of Rhode Island
Department of Psychology**



TEACHER CONSENT FORM FOR RESEARCH

My name is Monica Mabe, and I am asking for your participation in this study. I am working with Professor W. Grant Willis, a faculty member at URI.

Description of the project:

Recently, many observational tools have become available for use on phones or tablets. Last year, a study was conducted with one of these tools at the K-2 grade level for observing social skills. This year, I would like your help to conduct a study at the middle and high school level. One purpose of this project is to see if a behavioral observation tool can be used for observing how students engage in various social skills such as working in groups or following classroom expectations. The other purpose is to evaluate how social skills are measured through observation and through paper forms, called rating scales. This research project will begin in November and be completed in the school by February.

What will be done:

If you agree to participate, here is what will happen: A student from the University of Rhode Island (URI) will be assigned to observe students in your classroom and observe the students during their regular scheduled day. You will not be asked to change your classroom routine in any way. You will be asked to provide a daily schedule to the researcher so that the URI observer will know the best time to observe specific skills. You will also be asked to complete two social skills rating scales for participating students; the forms take 5-10 minutes to complete. Students will be asked to complete these forms as well with assistance from me or their clinician.

Risks or discomfort:

There are no risks or discomfort involved for you in this project. URI students will initially need assistance to identify specific students but should be of no further distraction to you or the class afterward.

Benefits of this study:

Although there may be no direct benefit to you for participating in this project, the school will benefit from the information that will be collected. The information from this project will help personnel to improve data collection practices in the future when assessing student behaviors. Results of this study will be made available for

viewing online through the ProQuest library search engine. A hard copy of results can also be viewed at the University of Rhode Island library after the completion of the study.

Confidentiality:

Your part in this study is confidential. The study is concerned with the students' skills and the measurement of those skills; the information you provide will not be linked back to you or used in any other capacity. Following data collection, all names will be removed and replaced with identification numbers so that identification cannot be traced back to a specific person.

Decision to quit at any time:

Your decision to participate will not affect your relationship with South Coast Educational Collaborative (SCEC). You also have the right to stop participating at any time.

Rights and Complaints:

If you are unhappy with the way this study is being conducted in your classroom, you may talk about your complaints with Professor W. Grant Willis (401) 874-4245 or with Graduate Student, Monica Mabe (435) 760-7213, both from URI. You may also talk to Charles Seekell (508) 336-8212 from SCEC regarding any concerns about this study. Please feel free to contact any of the individuals listed with further questions you may have about this research project. In addition, if you have questions about your rights as a research participant, you may contact the office of the Vice President for Research, 70 Lower College Road, Suite 2, University of Rhode Island, Kingston, Rhode Island, telephone: (401) 874-4328.

You have read this Permission Form. Your questions have been answered. Your signature on this form means that you understand the information and you agree to participate in this study. Thanks so much for your attention to this.

I agree to to participate in this study ___ Yes ___ No

Signature of Participant

Signature of Researcher

Typed/printed Name

Typed/printed name

Date

Date

Please sign both consent forms, keeping one for yourself

Literature Cited

- Arnold, d. H., Kupersmidt, J. B., Voegler-Lee, M. E., & Marshall, N. A. (2012). The association between preschool children's social functioning and emergent academic skills. *Early Childhood Research Quarterly, 27*, 376-386.
- Burns, M. K., & Coolong-Chaffin, M. (2006). Response to Intervention: The role of and effect on school psychology. *School Psychology Forum: Research In Practice, 1*, 3-15.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.
- Cappadocia, M. C., & Weiss, J. A. (2011). Review of social skills training groups for youth with asperger syndrome and high functioning autism. *Research in Autism Spectrum Disorders, 5*, 70-78.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.
- Cook, C. R., Gresham, F. M., Kern, L., Barreras, R. B., Thornton, S., & Crews, S. D. (2008). Social skills training for secondary students with emotional and/or behavioral disorders: A review and analysis of the meta-analytic literature. *Journal of Emotional and Behavioral Disorders, 16*, 131-144.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.
- Crowe, L. M., Beauchamp, M. H., Catroppa, C., & Anderson, V. (2011). Social function assessment tools for children and adolescents: A systematic review from 1988 to 2010. *Clinical Psychology Review, 31*, 767-785.
- Cummings, K. D., Kaminski, R. A., & Merrell, K. W. (2008). Advances in the assessment of social competence: Findings from a preliminary investigation of a general outcome measure for social behavior. *Psychology in the Schools, 45*, 930-946.
- Demaray, M. K., & Jenkins, L. N., (2011). Relations among academic enablers and academic achievement in children with and without high levels of parent-rated symptoms of inattention, impulsivity, and hyperactivity. *Psychology in the Schools, 48*, 573-586.
- Fisher, R. A. (1924) On a distribution yielding the error functions of several well known statistics. *Proceedings of the International Congress of Mathematics, Toronto, 2*, 805-813.

- Flanagan, D. P., Alfonso, V. C., & Dixon, S. G. (2014). Academic achievement batteries. In *Academic Assessment and Intervention*. Little, S. & Akin-Little, A. (Eds.), New York, NY: Routledge.
- Fletcher, J. M., & Vaughn, S. (2009). Response to Intervention: Preventing and remediating academic difficulties. *Child Development Perspectives*, 3, 30-37.
- Gresham, F. M. (2010). Evidence-based social skills interventions: Empirical foundations for instructional approaches. In Shinn, M. R., & Walker, H. M. (Eds.), *Interventions for achievement and behavior problems in a three tier model, including RTI* (337-362). Bethesda, MD: National Association of School Psychologists.
- 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, 364-379.
- Gresham, F. M., Cook, C. R., Crews, S. D., & Kern, L. (2004). Social skills training for children and youth with emotional and behavioral disorders: Validity considerations and future directions. *Behavioral Disorders*, 30, 32-46.
- Gresham, F. M., & Elliott, S. N. (2008). *Social Skills Improvement System: Rating scales*. Bloomington, MN: Pearson Assessments.
- Gresham, F. M., Sugai, G., & Horner, R. H. (2001). Interpreting outcomes of social skills training for students with high-incidence disabilities. *Exceptional Children*, 67, 331-344.
- Hintze, J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review*, 33, 258-270.
- Hoagwood, K., & Johnson, J. (2003). School psychology: A public health framework I. From evidence-based practices to evidence-based policies. *Journal of School Psychology*, 41, 3-21.
- Hosp, M. K., Hosp, J. L., & Howell, K. W. (2007). *The ABCs of CBM: A Practical Guide to Curriculum-Based Measurement*. New York, NY: Guilford Press.

- Jyoti, D. F., Frongillo, E. A., & Jones, S. J. (2005). Food insecurity affects school children's academic performance, weight gain, and social skills. *The Journal of Nutrition*, 135, 2831-2839.
- Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 33, 363-374.
- Luiselli, J. K., McCarty, J. C., Coniglio, J., Zorilla-Ramirez, C., & Putnam, R. F. (2005). Social skills assessment and intervention: Review and recommendations for school practitioners. *Journal of Applied School Psychology*, 21, 21-38.
- Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Test of Educational Achievement – Second Edition*. Circle Pines, MN: American Guidance Service.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Mabe, M. L. (2013). *Progress monitoring for prerequisite social skills: A generalizability study for measure development* (master's thesis). Retrieved from ProQuest.
- Matson, J. L., & Wilkins, J. (2009). Psychometric testing methods for children's social skills. *Research in Developmental Disabilities*, 30, 249-274.
- Mushquash, C. & O'Connor, B. P. (2006). SPSS, SAS, and MATLAB programs for generalizability theory analyses. *Behavior Research Methods*, 38, 542-547.
- Salvia, J., & Ysseldyke, J. E. (2004). *Assessment* (9th ed.). Princeton, NJ: Houghton Mifflin.
- Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, 44, 922-932.
- Stecker, P. M., Lembke, E. S., & Foegen, A. (2008). Using progress-monitoring data to improve instructional decision making. *Preventing School Failure*, 52, 48-52.
- Stichter, J. P., Herzon, M. J., O'Connor, K. V., & Schmidt, C. (2012). A preliminary examination of a general social outcome measure. *Assessment for Effective Intervention*, 38, 40-52.
- Wechsler, D. (2009). *Wechsler individual achievement test – Third edition*. San Antonio, TX: Psychological Corporation.

Welsh, M., Parke, R. D., Widaman, K., & O'Neil (2001). Linkages between children's social and academic competence: A longitudinal analysis. *Journal of School Psychology, 39*, 463-481.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *The Woodcock-Johnson III – Tests of Achievement*. Itasca, IL: Riverside.