2013

Progress Monitoring for Prerequisite Social Skills: A Generalizability Study For Measure Development

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

PROGRESS MONITORING FOR PREREQUISITE SOCIAL SKILLS: A GENERALIZABILITY STUDY FOR MEASURE DEVELOPMENT

BY

MONICA MABE

A THESIS SUBMITTED IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

IN

PSYCHOLOGY

UNIVERSITY OF RHODE ISLAND

2013

MASTER OF ARTS THESIS

OF

MONICA MABE

APPROVED:

Thesis Committee:

Major Professor W. Grant Willis

John Stevenson

Susan Brand

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2013

Abstract

Social skills are important components of social-emotional functioning that allow children to be successful in both the social and academic spheres of school. A review of social skills intervention literature is presented including issues influencing effectiveness. Concerns associated with assessing the effects of social skills interventions are discussed and a formative assessment tool for behavioral observation is presented. The use of generalizability theory is then examined as a psychometrically based approach to developing a measure for observing social skills. Four prerequisite social skill areas were identified: (a) Attending, (b), Raise Hand, (c) Hands to Self, and (d) Transition. Transition was divided into two components for a total of five observed skills. Students in an elementary school were observed during regular classroom activities on three different occasions for each skill. The reliability of this strategy was evaluated in order to assess the optimal number of occasions and observers needed in order to obtain adequate degrees of reliability. Results identified particular skills that can be observed more reliably than others, and what combination of parameters might lead to optimal reliability. Preliminary descriptive analyses suggest that ethnicity might play a role in student performance of specific skills. Results are discussed in terms of applied use for the measure in school settings for formative assessment and in terms of directions for future research.

Acknowledgement

This thesis could not have been completed without the contribution of many individuals: Thank you to my major professor, Dr. W. Grant Willis, for his feedback, support, and tireless editing. Thank you to my thesis committee members, Dr. John Stevenson and Dr. Susan Brand for their insight and support. Thank you to Dr. Brian O'Connor for his work on the G2 SPSS syntax, which allowed me to complete my data analyses. Thank you to Hui-Qing Yin for her statistical consultation. Finally, thank you to the school professionals who provided me with the setting for this research opportunity.

Table of Contents

Table of Contents (Continued)

Table of Contents (Continued)

List of Tables

List of Figures

PROGRESS MONITORING FOR PREREQUISITE SOCIAL SKILLS: A GENERALIZABILITY STUDY FOR MEASURE DEVELOPMENT

Chapter I: Introduction

Statement of the Problem

In schools, educators historically have been primarily concerned with promoting academic competence among students. The social-emotional health of students, however, frequently has been placed as secondary in importance. Within the last two decades, a growing emphasis has been seen in schools to promote adaptive social functioning for students. This has happened as more research has shown that social functioning plays an important role in students' abilities to thrive in school environments (Cappadocia & Weiss, 2011). Social skills may serve as academic enablers, facilitating academic achievement (Gresham, Cook, Crews, & Kern, 2004). Gresham (2010) provided one example of how social and academic domains can overlap. Here, a deficit in social skills could lead to behavioral and discipline problems in the classroom, which may make instruction and learning more difficult. Overall, students with poor social skills are at risk for internalized and externalized behavioral problems as well as poor academic achievement (Cook, Gresham, Kern, Barreras, Thornton, & Crews, 2008).

A review of the literature on social-skills interventions reveals patterns of ineffectiveness (i.e., lack of generalization of skills) and inconsistency in the method of effectiveness measurement (e.g., rating scales, behavioral observations, sociometric

ratings). In academic interventions in schools (e.g., for reading, writing, math), curriculum-based measurements (CBMs) are commonly used to assess an intervention's effectiveness. These kinds of measures provide a valid and efficient method for data collection and decision making (Burns & Coolong-Chaffin, 2006). Comparable measures, however, are not yet readily available for monitoring the progress of social-skills interventions (Gresham, Cook, Collins, Dart, Rasetshwane, Truelson, & Grant, 2010).

The proposed research aims to use an established methodological framework (i.e., generalizability theory) to evaluate the psychometric properties of a behavioral-measurement tool and its utility for observing basic skills that are prerequisite to social competence in a classroom setting (e.g., keeping hands to self, raising hand, and waiting to be called on). By using multiple observers in multiple settings to obtain observational data on multiple students' skills, it is hoped that a useful method and tool for progress monitoring of these skills can be established.

Critical Review of Literature

The following critical review focuses on aspects affecting the outcome, or overall effectiveness, of social skills interventions. Next, issues of social-skills measurement in contemporary research are considered. The usefulness of curriculum-based measurement and progress monitoring for social-skills interventions is also discussed. Finally, the psychometric framework of generalizability theory is presented to support the proposed social skills behavioral measure.

Effectiveness

As noted, school-based social-skills interventions often have been found to be lacking in effectiveness. One problem with effectiveness is that many interventions lack generalization instruction (Cappadocia & Weiss, 2011; Gresham, Sugai, & Horner, 2001). This issue often results in students being able to perform specified skills in the setting where instruction occurs but not in other settings (e.g., classroom, lunchroom, playground). The most common format for social-skills interventions is a pull-out (students removed from regular classroom), small group (with 4 to 6 students and 1 or 2 adults), averaging 2 to 3 hours per week (Gresham et al., 2001; Gresham et al., 2006). Individuals teaching social skills to students may not notice the lack of skill generalization because they typically observe and assess students only within the instructional, small-group setting. Assessment in this kind of setting alone may not allow the instructor to gauge student progress and modify instruction in order to improve student performance, or instruction effectiveness, in other settings.

A second issue influencing the effectiveness of interventions is the lack of attention to the nature of the skill deficit. Two kinds of social skills problems that

may be targeted for intervention are acquisition (have not learned skill) and performance (do not perform a previously learned skill) deficits (Gresham et al., 2004). Most interventions focus on acquisition deficits and instructors may not differentiate their instruction for those students who have performance deficits (Gresham et al., 2001). A failure to differentiate instruction to meet the specific skill deficit a student is exhibiting may make the intervention less effective. Proper assessment of social skills can identify what kind of skill deficit is present and can allow the instructor to provide instruction that fits the needs of particular students.

A third issue is the use of poor evaluation measures with an intervention. Beelman, Pfingsten, and Losel (1994) found that studies demonstrating the most effective intervention outcomes were likely to be focused on direct goal criteria (i.e., the performance of specific skills) versus broad constructs (e.g., social adjustment, problem solving).

Thus, major weaknesses in this area of research include (a) a lack of generalizability training imbedded within interventions, (b) a lack of attention to the kind of skill deficit displayed by the student, and (c) the intervention and associated assessment strategies. These weaknesses in this area of research are all related to issues of measurement.

Measurement Considerations

Several meta-analyses have been conducted evaluating the effectiveness of social-skills interventions. Within the last three decades, seven meta-analyses (Ang & Hughes, 2001; Beelmann et al., 1994; Durlak, Fuhrman, & Lampman, 1991; Losel & Beelmann, 2003; Quinn, Kavale, Mathur, Rutherford, & Forness, 1999; Schneider,

1992; Schneider & Bryne, 1985) have been published as well as a few reviews of the meta-analytic literature (Cook et al., 2008; Gresham et al., 2004; Gresham, et. al, 2001). Major findings from the meta-analytic literature are (a) many studies use outcome measures that are not directly linked to the skills taught, and (b) most studies use outcome-based evaluation rather than formative assessment. Failure to link outcome measures to specific skills being taught is likely to decrease the accuracy of assessment. Likewise, small changes in skill performance may go undetected when only outcome measures are used.

Table 1 (adapted from Ang & Hughes, 2001) presents the kinds of measures that have been used in social-skills intervention studies over the two decades prior to 2001. Most of the 41 studies listed include multiple outcome measures. Of these studies, 27 used behavior ratings, 15 used behavior observation, 19 used self-report measures, 16 used skills-acquisition measures, and 14 used sociometric measures termed, "social adjustment measures." Measures were categorized as behavior ratings if they were a teacher or parent behavior-rating form. They were categorized as behavioral observations if they were based on naturalistic observation. Measures that required the students to perform a skill in a role play, or simulated setting, or to use paper and pencil to demonstrate problem solving were categorized as skills acquisition. Measures assessing student feelings or perceptions, such as a self-esteem scale, were categorized as self-report measures. Sociometric measures, such as peer ratings of aggression or acceptance, or recidivism for problem behavior, were categorized as social adjustment.

Table 1. *Measures Used in Social Skills Intervention Studies**

Study	Behavior Rating	Behavior Observation	Self Report	Skill Acquisition	Social Adjust-ment
Arbuthnot & Gordon (1986)	X			X	X
Bierman, Miller, & Stabb (1987)	X	X			X
Camp, Blom, Heber, & Doorninck (1977)	X			X	X
Coats (1979)	X	X			
Dishion & Andrews (1995)	X	X			X
Dubow, Huesmann, & Eron (1987)	X				
Etscheidt (1999)	X	X			
Feindler, Ecton, Kingsley, & Dubey (1986)	X			X	X
Feindler, Marriott, & Iwata (1984)			X	X	X
Forman (1980)	X	X			
Greenleaf (1982)				X	
Guerra & Slaby (1990)	X		X	X	
Hollin & Courtney (1983)			X		
Hudley & Graham (1993)	X			X	X
Huey & Rank (1984)	X		X	X	

Study	Behavior Rating	Behavior Observation	Self Report	Skill Acquisition	Social Adjust-ment
Kazdin, Bass, Siegel, & Thomas (1989)	X	X			
Kazdin, Esveldt-Dawson, French, & Unis (1987)	X				
Kazdin, Siegel, & Bass (1992)	X	X	X		
Kendall, Reber, McLeer, Epps, & Ronan (1990)	X		X		
Kettlewell & Kausch (1983)		X	X	X	
Larkin & Thyer (1999)			X		X
Larson (1992)	X		X		X
Lee, Hallberg, & Hassard (1979)			X	X	X
Lochman (1985)		X			
Lochman (1992)		X	X	X	X
Lochman & Curry (1986)	X	X	X		
Lochman, Burch, Curry, & Lampron (1984)	X	X	X	X	X
Lochman, Coie, Underwood, & Terry (1993)	X		X		X
Lochman & Lampron (1986)		X			

Study	Behavior Rating	Behavior Observation	Self Report	Skill Acquisition	Social Adjust-ment
Lochman, Ampron, Gemmer, Harris, & Wyckoff (1989)	X	X	X		
Long & Sherer (1984)		X	X		
Niles (1986)	X			X	
Ollendick & Hersen (1979)			X	X	
Pepler, King, Craig, Byrd, & Bream (1995)	X				X
Prinz, Blechman, & Dumas (1994)	X	X			X
Spence & Marzillier (1981)	X		X	X	
Spence & Spence (1980)			X		
Tanner & Holliman (1988)	X	X			
Tremblay, Pagani-Kurtz, Masse, Vitaro, & Phil (1995)	X		X		X
Vaughn, Ridley, & Bullock (1984)				X	
Vitar&Tremblay (1994)	X		X	X	

*Adapted from Ang and Hughes (2001)

Ang and Hughes's (2001) categorization strategy provides a general perspective of the most common kinds of measures used in social-skills intervention studies. Not all meta-analytic studies have used the same categorization strategy, but

most have shown that similar kinds of measures were used. Beelman, Pfingsten, and Losel (1994) for example, conducted a meta-analysis of 49 studies. This meta-analysis showed almost identical categorization of the kinds of measures used in social-skills interventions. They found that 45 studies evaluated interventions using social-cognitive tests analogous to the skill-acquisition measures described by Ang and Hughes, 41 used parent or teacher reports, 38 used behavioral observations, 31 used self-reports, and 21 used sociometric or peer reports.

Many of the measures used in studies of social skills may be questionable for assessing social-skills outcomes. Durlak, Fuhrman, and Lampman (1991) conducted a meta-analysis and found that 58 of the studies used behavioral observation, 19 used peer rating/sociometric measures, 20 used achievement/intelligence tests, 78 used cognitive-performance measures, and 4 studies used objective performance measures (e.g., observation of a specifici task performance). In this meta-analysis, the achievement/intelligence tests and cognitive-performance measures were identified as being inappropriate for assessing outcome of social skills interventions. Quinn et al. (1999) also found a number of studies that used inappropriate measures. Of the 35 studies examined in this meta-analysis, 23 studies used sociometric measures, 28 used behavior ratings, 8 used personality tests, and 17 used academic achievement tests. These meta-analyses illustrate how inappropriate measures are frequently used to assess outcome in social-skills studies even though they have not been validated for this particular use.

There are two main problems with the use of the measures described here. First, many social-skills intervention studies have used irrelevant measures to evaluate

effectivenes. It is possible that a number of studies found that interventions were ineffective simply because they used a measure to evaluate an outcome that was unrelated to the content of the intervention. For example, it is unlikely that interventions designed to teach social skills would influence student achievement or cognitive ability directly. Evenso, Ang and Hughes (2001) identified 37 studies that used cognitive and achievement tests as an outcome measure. Quinn et al. (1999) noted that larger effects could be seen in evaluating interventions when instruction focused on teaching and measuring specific skills rather than interventions with a more global emphasis. These findings stress the importance of using an outcome measure that is directly linked to the skills that are taught within the social skills intervention when attempting to assess its effectiveness.

The second problem is that the measures described here have all used an outcome-evaluation format. These kinds of measures provide summative information in a global manner that may not indicate a student's standing on specific skill components. Not one of the studies used formative assessment approaches. Intervention instructors often fail to plan for generalization of skills to settings outside the intervention setting (Gresham, 2010), which adversely influences effectiveness. Instead, formative assessment could be used to assess student progress directly on target skills and then to identify specific intervention strategies, thereby potentially increasing the intervention's effectiveness. Related to the issue of formative assessment is the concept of change sensitivity. Change-sensitive measures allow one to observe small changes in performance over brief periods of time (Burns & Coolong-Chaffin, 2006; Gresham et al., 2010); the kinds of measures described here

seem more sensitive to stability than to change.

Behavioral observation, however, is one approach that lends itself to a change-sensitive format; it is also one of the most widely used assessment procedures by school psychologists (Hintz & Matthews, 2004). Traditionally, it has been used as an outcome measure to determine if a student can perform particular tasks subsequent to an intervention. Consideration should be given to the lack of reliability of direct observation that some studies have shown (Hintze & Matthews, 2004). It may be difficult to obtain a high level of reliability using direct observation with only a handful of observations (e.g., more observations increases reliability), but the time required to conduct observations may be less than that required for administering and scoring behavioral rating systems (and potentially more productive). Many of the behavioral rating systems that are used to measure social skills have over a hundred items for a teacher or parent to rate. Rating systems are typically used as a General Outcome Measure (GOM), which do not provide information about specific skill defecits and simply provide an overall general description of skills (Hosp, Hosp, & Howell, 2007). Moreover, this kind of GOM is not conducive to multiple administrations over brief periods, which would be necessary for a change-sensitive instrument.

As noted, GOMs typically have been used to determine if a student can perform particular tasks subsequent to an intervention. A change-sensitive model differs from a general-outcome model in that it uses multiple observations throughout an intervention to detect performance changes in particular skills. Directly observing student performance, with a curriculum-based measure, can be used for informing

instructional decisions (Hintze, Christ, & Methe, 2006). This method of measurement follows a formative-assessment, change-sensitive model and facilitates individualization of an intervention, thereby potentially improving the intervention's overall effectiveness.

Curriculum-Based Measurement and Progress Monitoring

A formative assessment method is often referred to as progress monitoring. Progress monitoring is an important aspect of a multi-tiered format of intervention used in schools that is often referred to as Response-to-Intervention (RTI). In the RTI process, students are given quality instruction in the classroom and their progress is checked, or monitored, in order to identify students who are struggling with various concepts; instruction is differentiated, or tailored, for those identified students and their progress is monitored on a more frequent basis (e.g., semi-weekly). Students who do not show adequate progress within a given time fram receive intensified instruction in particular areas and continue to have their progress monitored; these students may be considered for special-educational services (Bradley, Danielson, & Doolittle, 2005).

Fletcher and Vaughn (2009) reported that the primary goal of the RTI model is to improve academic and behavioral student outcomes. The major emphasis in schools, however, has been to use the RTI model for academic interventions rather than for social/behavioral interventions. For example, Fuchs and Fuchs (2009) note that the major goal of RTI is to prevent long-term and debilitating academic failure. There are many progress-monitoring materials, or curriculum-based measures (CBMs), for academic interventions; however, little attention has been given to

developing CBMs for behavioral interventions. Currently, there is no CBM for measuring short-term responses to social-skills interventions (Gresham et al., 2010).

The RTI process 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, Hosp, & Howell, 2007). Progress-monitoring provides the means of evaluating instruction and teacher decision making (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 differention of instruction.

Indeed, the most common use of CBM progress-monitoring is decision making. The proper use of CBM to monitor student progress and to inform instructional changes in response to data significantly improves student achievement (Stecker, Lembke, & Faegen, 2008). Hosp, Hosp, and Howell (2007) described four kinds of decisions that can be made using data from CBMs: (a) screening decisions, (b) progress-monitoring decisions, (c) diagnostic decisions, and (d) outcome decisions.

In school settings, where social interactions are abundant, it is clear that there is a need for effective social-skills interventions. Schools using an RTI format are likely to require teachers to use CBMs to monitor the progress of their students and to adjust instruction accordingly. The use of CBMS, however, has been largely neglected in the area of behavioral interventions and there are no CBMs currently available for dependably measuring student response to short-term interventions in the area of social skills (Gresham et al., 2010). Brief rating scales have been developed

(Gresham & Elliott, 2008; Gresham et al., 2010) in order to improve the ability to progress monitor social skills, but have been developed in the style of traditional rating scales and bring with them all the associated difficulties that were previously discussed. The implementation of adequate progress monitoring for social skills interventions would likely increase the intervention's effectiveness by allowing instructors to assess and to monitor student performance over brief periods, as well as to adjust instruction as needed based on student performance.

Measurement-Theory Perspectives

A major perspective in psychometric assessment is classical test theory (CTT). 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 this theory is that error is randomly distributed and comes from sources unrelated to true differences in the assessed trait.

Generalizability theory (GT) is an extension of CTT that includes multiple sources of measurement error and that can be used to assess the dependability of behavioral measurements. Shavelson, Webb, and Rowley (1989) described the multiple ways that GT extends CTT: (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. GT is useful for assessing the reliability of CBMs, for example, because it accounts for error attributed both to multiple observers and to multiple settings. CTT is less than optimal for this

kind of assessment and, if used, might result in a lower reliability statistic than is desirable for efficient decision-making purposes. 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).

As noted, GT differentiates between two phases of a study: (a) generalizability studies and (b) decision studies. These two phases work together to optimize the reliability of a measure. The generalizability-study phase estimates the magnitude of potential sources of error whereas the decision-study phase uses this information to help to design a strategy that minimizes error for a specific purpose (Shavelson et al., 1989). In other words, the decision study allows one to estimate how adjustments to sources of error may affect reliability. For example, in a study assessing a behavioral-observation tool, a generalizability analysis could estimate the amount of error associated with the observer, the skills being observed, and the number of observations; a decision analysis could estimate what changes in one or several of these parameters might best improve the measure's reliability.

Purpose of the Study

Instructors of social-skills interventions need a better method of evaluating intervention effectiveness and student progress. An effective change-sensitive progress-monitoring tool is necessary to help instructors gauge student progress on specific skills, differentiate instruction appropriately, and improve the overall effectiveness of social-skills interventions. This study used 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

implementation of this tool in social-skills interventions will be able to improve program effectiveness and student outcomes.

Chapter II: Method

Participants

Participants were 31 elementary-school students from intact classrooms in kindergarten through second grade (ages 5 to 8 yrs., M = 6.7 yrs., SD = .9 yrs., Median = 7 yrs) attending a charter school in the northeastern part of the United States. Sample size was selected given considerations for the statistical analyses that were conducted. Grade levels were chosen by administrative staff at the school given expressed teacher interest and accessibility of the classrooms to outside observers. The majority of students in this school were from African American and Hispanic ethnic backgrounds and from families of low socio-economic status (SES). SES was estimated by participation in the school's lunch program: Students who qualified for a 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. About half (i.e., 51%) of the sample qualified for free lunch (low SES), 13% for reduced-price lunch (medium to low SES), and 36% paid the full price for lunch (medium to above SES). The characteristics of the sample are presented in Table 2.

Table 2. *Sample Characteristics*

Characteristic		*n*	%
Gender	Male	14	45%
	Female	17	55%
Ethnicity	African American	18	58%
	Hispanic	6	19%
	White	6	19%
	Multiple	1	3%

Characteristic		*n*	%
SES	Low	16	51%
	Low to middle	4	13%
	Middle to above	11	36%
Age	5 years	4	13%
	6 years	7	23%
	7 years	15	48%
	8 years	5	16%
Grade level	Kindergarten	9	29%
	First	11	35%
	Second	11	35%

A detailed description of school-wide student demographics appears in Appendix A as well as a brief review of multicultural considerations in this study.

Measures

Measurement instrument. The measurement instrument used in the current study is named *Metryx.* This observational tool was developed by Stephanie Castilla and Shawn Rubin at the participating school to supplement traditional observation techniques and to provide a technological option for obtaining observational data. Rubin is a former elementary educator and Castilla is an industrial designer; they worked together to build a technology that could replace traditional pen and paper options that teachers had for recording student data. The goal was to build a mobile formative assessment platform that would allow all teachers to work with data in real time.

Metryx uses iPad technology for tracking classroom academic data in an RTI format. It was founded on the belief that the best teaching is personalized; high-achieving students should receive acccelerated instruction, and students who struggle should receive targeted instruction. Metryx was designed to provide an effective and efficient tool to collect, to analyze, and to differentiate based on formative data to

guide instruction. Because Metryx received positive reviews from teachers using it in their classrooms, specialists at the school began to wonder how else Metryx could be used. The focus then turned to how Metryx could be used to collect observational data on social skills and to inform intervention instruction.

Metryx was designed to be used by an observer, who selects a desired skill set such as engaging in conversation and is given a list of the skill components underlying that concept (e.g., verbal initiation, eye contact, etc.). The observer taps an iPad under the designated skill being observed to indicate that a target skill was observed as successfully or unsuccessfully completed. Metryx is able to provide charts of progress instantly based on current and past observations in various social skills. The collected information can be used in the future for decision-making purposes about a student's progress and educational needs as well as to provide both ipsative and normative comparisons. In other words, Metryx provides feedback about an individual student's progress toward personal goals as well as progress compared to peers.

Dependent variables. Dependent variables in the current study were ratings of successful completion in four specified social skill areas: (a) attending to lesson, (b) keeping hands to self, (c) raising hand and waiting to be called, and (d) transitioning. "Attending to lesson" was defined as being actively or passively involved in the lesson (i.e., being "on-task"). Examples include looking at the teacher during instructional periods or participating in specified tasks; nonexamples include participating in an activity that is non-compliant with the lesson, talking to others during instructional periods, and being otherwise engaged during instructional periods. "Keeping hands to self" was defined as keeping one's hands within personal space and out of others'

space. Nonexamples include touching others and touching others' property without invitation. "Raising hand and waiting to be called" was defined as a student raising the hand in class and waiting to be called on before speaking. Nonexamples include speaking out of turn while raising the hand or speaking out without raising the hand. "Transitioning" was divided into two parts for observation purposes. The first part of Transitioning that was observed was "Transition: Quiet." This was defined as the completion of a transition task (specified by teacher) quietly without disrupting other students. The second part of Transitioning was "Transition: Follow Directions." This was defined as the completion of a transition task (specified by teacher) quickly and well. Examples include completing all components of a transition task specified by the teacher within a brief time period without additional prompting.

These classroom behaviors were chosen because they are easily observable and serve as precursors to the social skills that are taught in intervention groups for these grades. Skills such as attending, listening, staying on task, and following directions have been shown to affect students' readiness to learn and can affect individual and classroom behaviors linked to academic and social success (Villares, Brigman, & Peluso, 2008). Additionally, teachers at the participating school identified these particular skills as essential for success in the classroom environment.

Both parts of "Transitioning" as well as "Raising hand and waiting to be called" were scored on a rubric continuum ranging from 1 through 5. A rating of 1 indicated that a student was not successful in the transition or raising hand and waiting to be called, 2 indicated the student was somewhat successful, 3 indicated that the student succeeded in completing half of the criteria, 4 indicated that the student was

mostly successful, and 5 indicated that the student completed criteria nearly flawlessly. Momentary time sampling was used to record observations for skills "Attending" and "Keeping hands to self." Momentary time sampling required that a student be observed at the end of each 30-second interval to determine if that student was engaged in the specified behavior at that given moment. These skills were scored as "no" (not observed) or "yes" (observed) during each interval of the observation; a total of 30 intervals occured during the observation and an overall percentage completed was calculated for each observational occasion. These percentages were then converted to the same rubric continuum as "Transitioning" and "Raising hand and waiting to be called," with values of 0% to 19% scored as 1, 20% to 39% as 2, 40% to 59% as 3, 60% to 79% as 4, and 80% to100% as 5.

Procedures

Informed Consent/Assent

Informed consent was obtained from parents for student participation in the study and assent was obtained from the students; informed parent consent and student assent were the only inclusion criteria. English and Spanish versions of the informed consent form (Appendix B) were mailed to parents. Consent forms were sent to 140 homes and 22% (33 parents) were signed and returned with permission to participate. Student assent forms (Appendix C) were read aloud to students; they were asked to write their name on the format and to mark an X next to a "yes" or "no" for their decision to participate; two of the students for whom parent consent was obtained did not assent to participate. Observational 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 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

Three Caucasian females and one Caucasian male enrolled in a psychology undergraduate program served as observers for course credit and were trained in the use of Metryx. Each observer attended two one-hour training sessions conducted by the researcher as well as an additional one-hour training session conducted by one of the developers of Metryx. During these training sessions, observers discussed operational definitions of observational behaviors and were trained in the use of momentary time sampling, partial-interval recording, and frequency recording. Observational skills were practiced while observing video recordings of children in a classroom. Observers also received training in what is considered to be appropriate classroom demeanor and how to use timing devices properly during observational periods. Additionally, observers practiced using Metryx before entering classrooms and conducted practice observations in each classroom using Metryx before official data collection began. The researcher watched each of the observers conduct their first observation in order to assess their proficiency with Metryx, timing devices, and classroom demeanor. All observers 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 schools.

Direct observation. Students were observed during the naturally occurring day in the classroom and transition periods. Dates, times, and locations of observations as

well as the subject matter being studied during an observation were recorded. The skills, "Attending" (momentary time sampling) and both parts of "Transitioning" (rubric scoring), were observed on the same occasions as one skill pair and the skills, "Keeping hands to self" (momentary time sampling) and "Raises Hand" (rubric scoring), were observed on a separate occasion as the second skill pair. Thus, each pair of skills had one skill area observed with momentary time sampling and one skill area observed through rubric scoring. The rubric scoring systems required observers to document whena behavior occurred, whereas the momentary time sampling procedures required observers to observe students across a 15-minute period. Thus, the researcher paired them together in order to maximize the productivity of time spent in observation.

The undergraduate observers were randomly assigned to students in consideration of the amount of time they were able to devote to the research. Each participating student was observed on three occasions on each skill-area pair; each observation was 15-minutes long, divided into thirty-second intervals. Thus, each participating student was observed for 15 minutes in the classroom on 6 separate occasions for a total of 90 minutes. Table 3 illustrates this observational matrix.

Table 3. *Observation Matrix*

Observer	Occasion	Skill			
		Pair 1		Pair 2	
		Attend	Transition	Raise Hand	Hands to Self
A	1	$S_{1\text{-}8}$	$S_{1\text{-}8}$	$S_{1\text{-}8}$	$S_{1\text{-}8}$
	2	$S_{1\text{-}8}$	$S_{1\text{-}8}$	$S_{1\text{-}8}$	$S_{1\text{-}8}$
	3	$S_{1\text{-}8}$	$S_{1\text{-}8}$	$S_{1\text{-}8}$	$S_{1\text{-}8}$
	1	$S_{9\text{-}16}$	$S_{9\text{-}16}$	$S_{9\text{-}16}$	$S_{9\text{-}16}$

Observer	Occasion	Skill			
		Pair 1		Pair 2	
		Attend	Transition	Raise Hand	Hands to Self
B	2	$S_{9\text{-}16}$	$S_{9\text{-}16}$	$S_{9\text{-}16}$	$S_{9\text{-}16}$
	3	$S_{9\text{-}16}$	$S_{9\text{-}16}$	$S_{9\text{-}16}$	$S_{9\text{-}16}$
C	1	$S_{17\text{-}23}$	$S_{17\text{-}23}$	$S_{17\text{-}23}$	$S_{17\text{-}23}$
	2	$S_{17\text{-}23}$	$S_{17\text{-}23}$	$S_{17\text{-}23}$	$S_{17\text{-}23}$
	3	$S_{17\text{-}23}$	$S_{17\text{-}23}$	$S_{17\text{-}23}$	$S_{17\text{-}23}$
D	1	$S_{24\text{-}31}$	$S_{24\text{-}31}$	$S_{24\text{-}31}$	$S_{24\text{-}31}$
	2	$S_{24\text{-}31}$	$S_{24\text{-}31}$	$S_{24\text{-}31}$	$S_{24\text{-}31}$
	3	$S_{24\text{-}31}$	$S_{24\text{-}31}$	$S_{24\text{-}31}$	$S_{24\text{-}31}$

The undergraduate observers spent between 3 and 6 hours each week observing the participating students in their assigned classrooms. Observations were collected for 3 months from March through May of 2012 (with a total of one week off for school break in March) until all observations were completed.

Additionally, inter-rater observations were conducted for 20 randomly selected students on the third occasion for each skill. Secondary observers were randomly assigned to students and conducted an observation 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.

Design

For practical reasons, the study was designed with students nested in observers (i.e., these variables were not completely crossed). Students were not included as a separate facet because this would have required each student to be observed on six separate occasions by each observer (for a total of 30 observations per student, or N = 930 observations). Thus, the generalizability study was conceptualized as a three-

facet, partially nested design with occasions (3 levels) and skills (5 levels) as crossed factors, and students (N = 31) nested within observers (4 levels). Other potential sources of variation that were not assessed in this study included teacher, grade-level (or age), and activity or subject matter completed during the observations, among others. The dependent variable was the observational outcome, or score, on a 1 to 5 point Likert-type scale rubric for each of the five skills on each of three occasions for each of the 31 students (N = 465 measures). This design allowed for an estimation of variance components for (a) students nested within observers; (b) skills; (c) occasions; (d) 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.

Chapter III: Results

Four sets of analyses were conducted: (a) G-study analyses, (b) D-study analyses, (c) Kappa analyses of inter-observer agreement, and (d) MANOVA analyses of demographic characteristics.

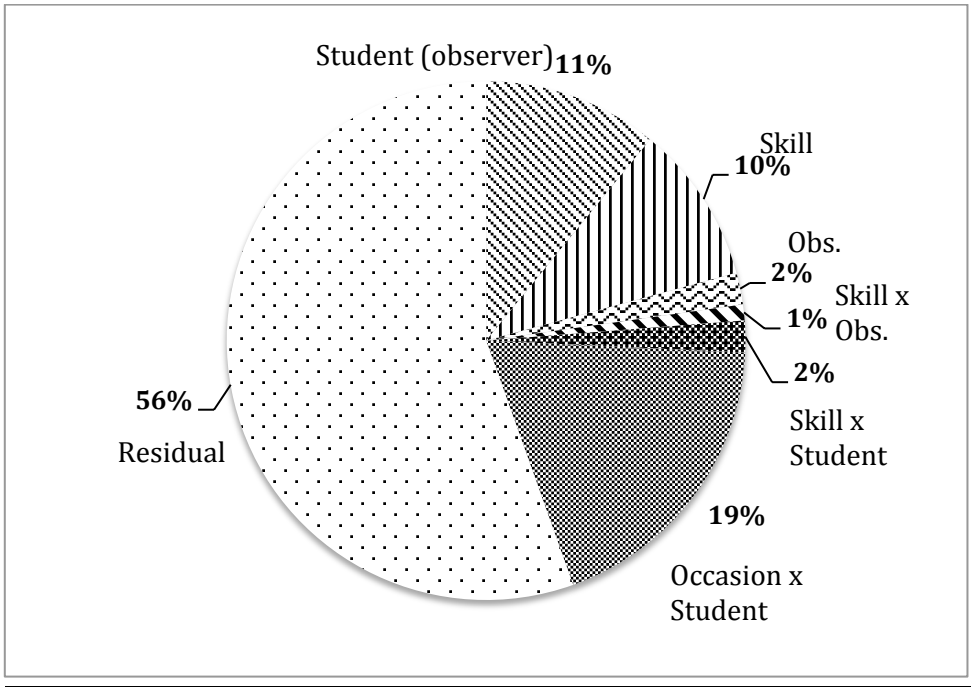
Generalizability and Decision Studies

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 4, 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 4: *G-Study Results*

Source of Variation	Variance Component	Proportion of Variance
Student (Observer)	.072	.105
Skill	.070	.102
Occasion	.000	.000
Observer	.011	.016
Skill × Occasion	.002	.002
Skill × Observer	.006	.008
Occasion × Observer	.000	.000
Skill × Student (Observer)	.015	.022
Occasion × Student (Observer)	.130	.190
Residual	.380	.555
Total	--	1.00

Figure 1: *Percents of Variance Explained*



The overall, relative G-Coefficient (used for decisions based on the relative standing of comparison to others) of the measure was .80. The residual term accounted for the greatest portion of variance (i.e., 56%). Students (i.e., the object of measurement, accounting for nesting within observer), however, only accounted for approximately 11% of the variance, and Skill accounted for 10%.

Skills Analyses

In order to determine if different skills were associated with different reliability estimates, separate G-studies were conducted according to the skill assessed. A series of five two-facet, partially nested G-studies was conducted, using the G2.sps program previously described, to examine the data separately for each of the five skills assessed and to identify any unique features specific to those skills. Variance components and proportions of variance accounted for were calculated for Students nested within Observers, Occasions, Observers, the Observer-by-Occasion interaction, and a

residual term that included the three-way interaction combined with error. These G-study results appear in Table 5.

Table 5. *Variance Components/Proportions of Variance and Relative G-coefficients for Individual Skills*

Source of Variation	Skills				
	Attending	Raise Hand	Hands to Self	Transition: Quiet	Transition: Follow Directions
Student (Observer)	.047/.110	.379/.366	.000/.000	.074/.086	.056/.067
Occasion	.000/.000	.006/.006	.000/.000	.015/.017	.000/.000
Observer	.008/.019	.018/.017	.000/.000	.059/.069	.000/.000
Ocassion × Observer	.000/.000	.000/.000	.000/.000	.000/.000	.000/.000
Residual	.371/.871	.633/.611	.000/.000	.716/.829	.769/.933
G-coefficient	.602	.878	1.000	.553	.465

Next, five separate D-studies were conducted in order to estimate how varying levels of facets might affect the reliability of each of those facets. These D-study results are presented in Table 6.

Table 6. *D-Study Results for Skills*

Skills	Occasions	Observers			
		1	2	6	8
Attending	1	0.112	0.202	0.431	0.502
	2	0.202	0.335	0.602	0.669
	4	0.335	0.502	0.752	0.802
	5	0.387	0.558	0.791	0.835
Raise Hand	1	0.374	0.545	0.782	0.827
	2	0.545	0.705	0.878	0.905
	4	0.705	0.827	0.935	0.950
	5	0.750	0.857	0.947	0.960
Hands to Self	1	0.000	0.000	0.000	0.000
	2	0.000	0.000	0.000	0.000

Skills	Occasions	Observers			
		1	2	6	8
Hands to Self	4	0.000	0.000	0.000	0.000
	5	0.000	0.000	0.000	0.000
Transition: Quiet	1	0.094	0.171	0.339	0.398
	2	0.171	0.292	0.490	0.554
	4	0.292	0.452	0.630	0.688
	5	0.340	0.508	0.668	0.723
Transition: Follow Directions	1	0.067	0.126	0.303	0.367
	2	0.126	0.224	0.465	0.537
	4	0.224	0.367	0.635	0.698
	5	0.266	0.420	0.685	0.743

Figures 2 through 5 provide graphic illustrations of these relative G-coefficients for Attending, Raise Hand, Transition: Quiet, and Transition: Follow Directions, respectively. (Hands to Self is not included because there was no variability in the ratings for any student on any occasion for this skill.)

Figure 2. *D-Study Results for Attending*

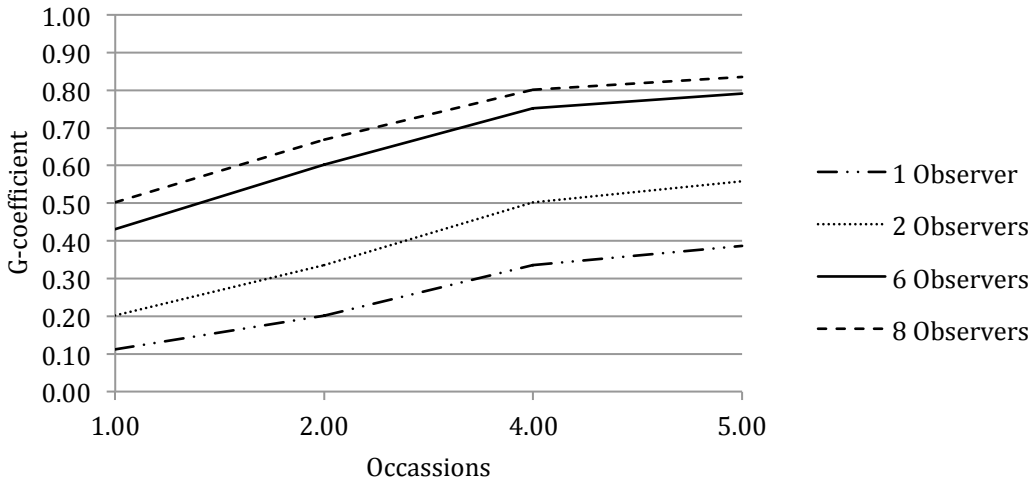


Figure 3. *D-Study Results for Raise Hand*

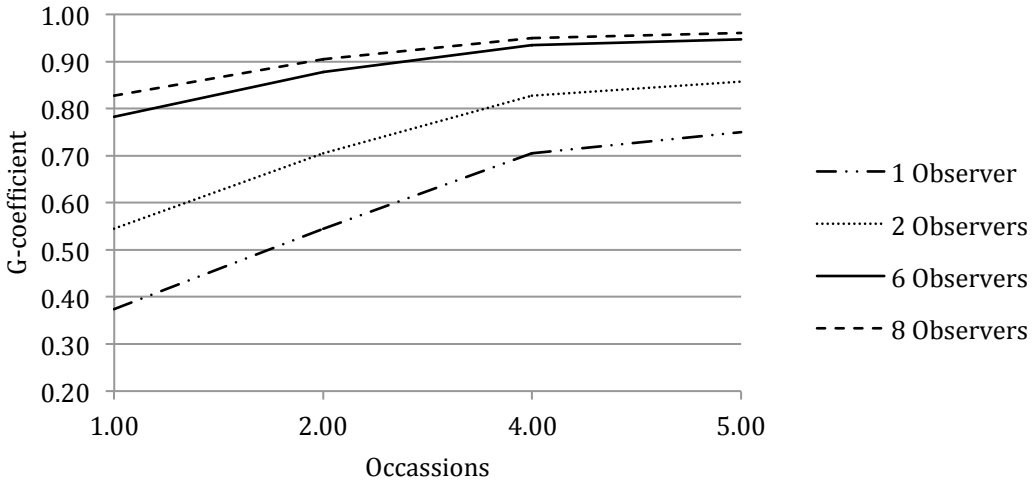


Figure 4. *D-Study Results for Transition-Quiet*

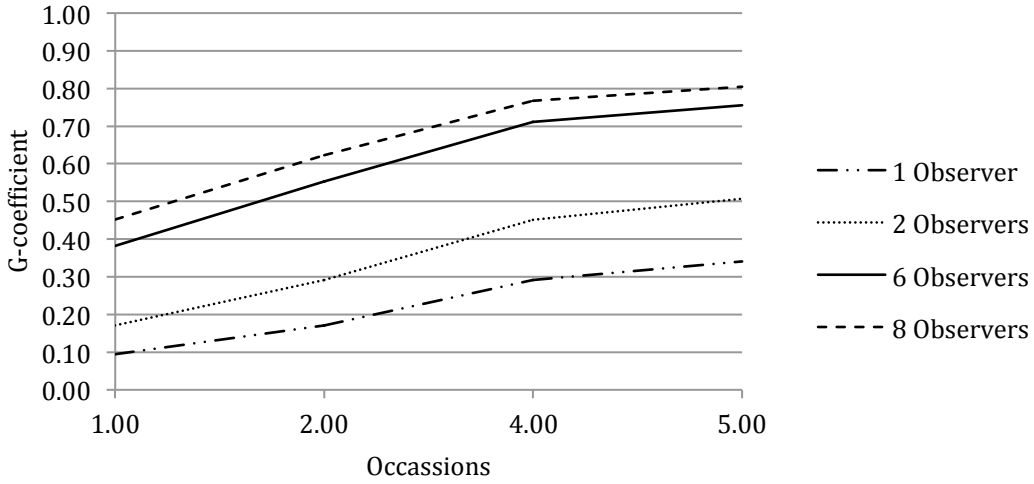
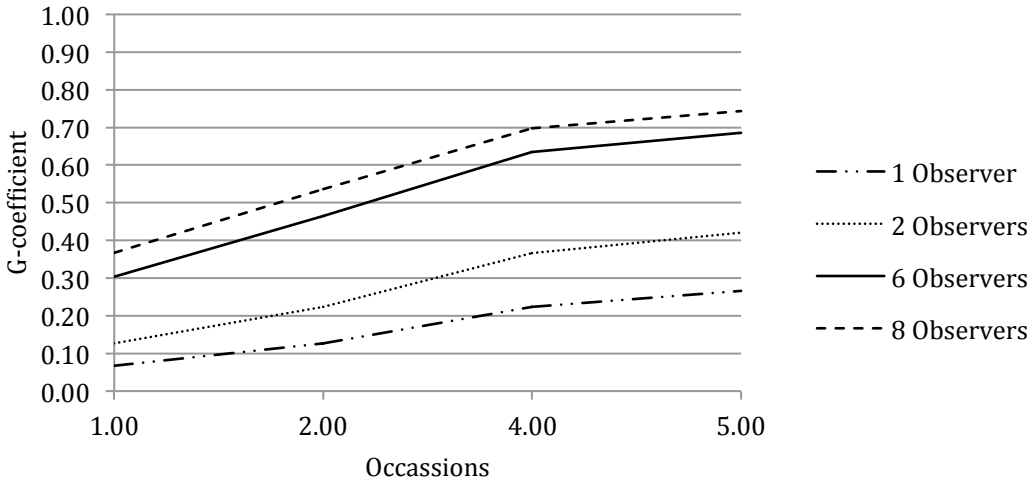


Figure 5. *D-Study Results for Transition-Follow Directions*



Attending. As previously described, this was defined as the level to which a student was paying attention to or participating in the lesson or activity at the time of observation. The largest proportion of variance (i.e., 87%) for this skill was accounted for by the residual term. The second largest contributor to the variance was the object of measurement, that is, students nested within observers. Here, 11% of the variance for Attending could be accounted for by individual students, taking into account that observers were assigned to specific groups of students for observations. The observer facet accounted for only about 2% of the variance for Attending. These results indicate that students varied in their ability to attend to the lesson or activity; it was the students' levels of skill, and not the rating style of the observer, that accounted for most of the variance.

The relative G-coefficient, which describes the universal reliability of the measure, for Attending was .602. This is a below adequate level of reliability for a behavioral measure and indicates that nearly 40% of the variance associated with the

measure was due to error. According to the D-study results, a combination of 5 occasions with 8 observers would be expected to improve reliability to approximately .835. This indicates that a good level of reliability for Attending can be obtained by adding 2 occasions and 3 observers to the present design.

Raise Hand. As previously described, this was defined as raising the hand and waiting to be called on while not speaking out of turn. The largest proportion of variance (i.e., 61%) for this skill was accounted for by the residual term. The second largest contributor to the variance was the object of measurement, that is, students nested within observers. Here, 37% of the variance for Raising Hand could be accounted for by individual students, taking into account that observers were assigned to specific groups of students for observations. The observer facet accounted for only about 2% of the variance for Raising Hand. These results indicate that students varied in their ability to raise their hand and wait quietly to be called on; it was the students' levels of skill, and not the rating style of the observer, that accounted for most of the variance.

The relative G-coefficient for Raise Hand was .878. This is a good level of reliability for a behavioral measure and indicates that only about 13% of the variance associated with the measure was due to error. According to the D-study results, a combination of 4 occasions with 6 observers would be expected to improve reliability to approximately .94. A combination of 5 occasions and 8 observers predicted the optimal level of reliability (i.e., .96), but the difference between this and the previous combination is negligible and would require much greater commitment of time and resources. Thus, although the obtained reliability of .878 was adequate for measuring

Raise Hand, it could be increased to .94 with the reasonable addition of 1 occasion and observer to the present design.

Hands to Self. As previously described, this was defined as the student keeping hands within personal space and out of others' space. Variance components for facets, proportion of variance accounted for by facets, and an overall G-coefficient could not be calculated for Hands to Self because the data were completely homogenous. That is, all ratings for students on this skill were exactly the same, which resulted in a lack of variance for this skill. These results indicate that the Hands to Self skill, as defined in the present study, was not well-suited to this type of behavioral observation.

Transition: Quiet. As previously described, this was defined as the quiet completion of a transition task without disrupting other students. The largest proportion of variance (i.e., 83%) for this skill was accounted for by the residual term. The second largest contributor to the variance was the object of measurement, that is, students. Here, 9% of the variance for Transition: Quiet could be accounted for by students. Observers, accounted for only about 7% of the variance. These results indicate that the rating style of the observer contributed almost as much to the variance in scores for Transition: Quiet as that of the performance of skill by the students.

The relative G-coefficient for Transition: Quiet was .553. This is a below adequate level of reliability for a behavioral measure and indicates nearly 45% of the variance associated with the measure was due to error. According to the D-study results, a combination of 5 occasions with 8 observers would be expected to improve reliability to approximately .723. This indicates that an more than 2 occasions and 3

observers would need to be added in order to obtain an adequate level of reliability for Transition: Quiet.

Transition: Follow Directions. As previously described, this was defined as the completion of a transition task quickly and well. The largest proportion of variance (i.e., 93%) for this skill was accounted for by the residual term. The second largest contributor to the variance was the object of measurement, that is, students nested within observers. Here, about 7% of the variance for Transition: Follow Directions could be accounted for by individual students, taking into account that observers were assigned to specific groups of students for observations. The observer facet accounted for virtually none of the variance for Transition: Follow Directions. These results indicate that students varied in their ability to follow directions from the teacher on transition tasks; it was the students' levels of skill, and not the rating style of the observer, that accounted for most of the variance.

The relative G-coefficient for Transition: Follow Directions was .465. This is a below adequate level of reliability for a behavioral measure and indicates that nearly 54% of the variance associated with the measure was due to error. According to the D-study results, a combination of 5 occasions with 8 observers would be expected to improve reliability to approximately .743. This indicates that an more than 2 occasions and 3 observers would need to be added to the present design in order to obtain an adequate level of reliability for Transition-Follow Directions.

Interobserver Agreement

Interobserver agreement was calculated for all observer pairs across skills. A randomly selected group of 20 students was assigned to each secondary observer

simultaneously with the student's previously assigned observer on the third occasion for each skill. In other words, 20 students were observed on each of the skills by two observers on the third occasion of observation. 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. Kappa values from 0.41 to 0.60 have been categorized as moderate, and values above 0.60 as substantial (Landis & Koch, 1977). Table 7 displays the level of agreement for primary and secondary observers across skills.

Table 7. *Interobserver Agreement Kappa Values*

Skills	Kappa
Combined	.364
Attending	.418*
Raise Hand	.438*
Hands to Self	N/A
Transition: Quiet	.161
Transition: Follow Directions	.246

*Moderate agreement. N/A: Could not be calculated

Across all skills combined, primary and secondary observers displayed agreement slightly higher than chance. Moderate agreement was found between primary and secondary observers for Attending and Raise Hand. Primary and secondary observers did not display substantial agreement on any of the observed skills. Level of agreement could not be calculated for Hands to Self because the ratings for this skill were homogenous. Results from the previously described G and D studies indicated that rating style of the observers influenced scores on at least two of the skills (i.e., Hands to self and Transition-Quiet). Assessment of potential

differences of ratings for different skills was not calculated as the number of ratings for individual observer pairs for each skill were so small that interpretation would not be meaningful.

Descriptive Analyses

A series of four, multivariate analysis of variance (MANOVA) tests were conducted in order to assess any score differences based on demographic categories. Dependent variables included scores on each of the five skills (including the two subskills for Transition) that were observed. Age was treated as a categorical variable (5, 6, 7, and 8 years), Ethnicity included four groups (African American, Hispanic, White, and Multi-ethnic), and SES included three groups (low, low to medium, and medium and above, as previously described). These results are considered catiously as exploratory because of inadequate statistical power owing to low sample sizes. Table 8 provides MANOVA results for demographic factors.

Table 8. *Results of MANOVAs for Demographic Factors*

Factor	Wilks' λ	*F*	*df*	Error *df*	*p*	η^2
Age	.639	.858	12	55.852	.592	.139
Sex	.961	.230	4	23.000	.918	.039
Ethnicity	.448	2.593	8	42.000	.021	.331
SES	.554	1.886	8	42.000	.100	.257

The only significant MANOVA was for Ethnicity, which showed a skewed distribution with disproportionately more participants who identified as Hispanic (*n* =18) than African American (*n* = 6) or White (*n* = 6). Participants identifying as Multi-ethnic were not represented in this analysis as there was such small representation (*n* = 1). The multivariate effect size for Ethnicity was substantial (η^2 =

.331), and indicated that ethnicity appears to influence student scores on specific skills. Table 9 presents follow-up analysis of variance tests (ANOVA) for Ethnicity according to each of the five skills.

Table 9. *Results of ANOVAs for Ethnicity*

Dependent Variable	SS	*df*	MS	*F*	*p*	η^2
Attend	1.460	2	.730	2.481	.105	.171
Raise Hand	1.658	2	.829	.944	.403	.073
Hands to Self	0	2	0	--	--	--
Transition: Quiet	6.401	2	3.200	3.818	.036	.241
Transition: Directions	3.231	2	1.616	1.317	.287	.099

The only significant ($p < .05$) ANOVA was for Transition: Quiet, which showed a substantial effect size (($\eta^2 = .241$). Followup Tukey tests show a difference between scores for Transition: Quiet ($p < .05$), with students who identfy themselves as White being observed to have completed transition activities quietly without disrupting other students more frequently than students who identify themselves as Hispanic.

Chapter IV: Discussion

Social skills are important for student social and academic success in school (Cappadocia & Weiss, 2011; Cook, et al., 2008; Gresham, Cook, Crews, & Kern, 2004). Prerequisite social skills, like those observed in this study, are essential to student success in the classroom and enable a child to function appropriately in a school environment (Villares, Brigman, & Peluso, 2008). This study used G theory to develop a measure for behavioral observation, specifically for the purpose of progress monitoring social skills. G theory was chosen for this study because of the many benefits it has over the traditional approach of CTT. In G theory, multiple facets of a research design can be examined in consideration of reliability and residual error, as compared to CTT which only considers a single main effect and assumes all other variance to be random error. G theory also allows for the prediction of reliability of measurement given different levels of facets than those used in the original design of the study. For example, one could estimate how reliable a measure would be if there were fewer or more observers or occasions; one could estimate the least amount of resources needed in order to maintain a good level of reliability. This aspect of G theory is especially useful in schools where resources are limited and information-gathering needs are high.

Psychometric Findings

The present study used G theory to examine the reliability of an observational tool to observe student performance of prerequisite social skills, with student nested within observer as the object of measurement and occasion, skill, and observer as facets. In a nested design, 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 students were observed on one skill while other students were observed on another skill. In this study, students were nested within observers. Each observer was assigned a particular number of students to observe; students are nested within observers because each observer did not observe every student on every occasion.

Relative G-coefficients were reported as a measure of overall reliability for the measure. Relative coefficients were also reported for all decision studies for relative interpretations rather than absolute interpretations. As stated previously, relative decisions, or interpretations, are those concerning an individual's performance relative to others. Absolute decisions, or interpretations, are those concerning an individual's performance compared to a specific criterion regardless of other's performance. Relative decisions could be used with the present data to screen students for social-skill performance in order to form intervention groups for students with similar levels of need. The relative G-coefficient is analogous to the reliability coefficient in classical test theory and is a more accurate indicator of reliability than the absolute Phi-coefficient of dependability (Shavelson & Webb, pg 93). Thus, relative G-coefficients were reported for the purpose of relative interpretations in the present study.

The relative G-coefficient was .80, an acceptable level of reliability for a behavioral measure. The largest proportion of variance was accounted for by residual error (i.e., 56%), which includes all 3-way interactions between facets that cannot be statistically partialled out. Student performance accounted for about 11%. The

Occasion by Student interaction accounted for the second-largest proportion of variance (19%) and Skill accounted for about 10%. The 10% of variance accounted for by Skill indicates that the type of skill being observed may have had an impact on observer ratings of student performance. The observed skills may not have been homogenous enough to be grouped together in the manner used in this study. In order to examine the Skill facet in more depth, five additional nested design G-studies were conducted (one for each skill) with Students nested within Observers as the object of measurement with Occasions and Observers as secondary facets.

The G-studies conducted on each skill resulted in varying degrees of reliability. One skill had an acceptable level of reliability (i.e., Raise Hand), but many of the skills had levels of reliability that were unacceptable (i.e., Attending, Transition-Quiet, Transition-Follow Directions). The skills chosen for observation may not have been homogenous, or assessed similar underlying skills that made their grouping conceptually similar. Lacking homogeneity is one explanation for the varying levels of reliability for each skill. In addition, results indicated that the rating style of observers may have influenced scores differently for each skill (e.g., each observer rated skills differently from one another). A discussion of each skill's G and D study results as well as interobserver effects follows.

Attending

Attending had a relative G-coefficient of .60; D-study results indicated that the reliability could be increased to .83 with a combination of five occasions and eight raters (this is an addition of two occasions and three raters to the present design). The additional resources required to obtain this level of reliability might be out of reach for

a typical school setting. Progress-monitoring procedures support the monitoring of target skills twice a week; this would allow for at least 5 observations to be completed before making decisions regarding student progress on Attending (e.g., after three weeks of observations). The number of observations required for adequate reliability for Attending appears to be within reach for a typical school, but the number of observers required is likely out of reach. There are three likely candidates to conduct observations in a typical school setting: School psychologists, social workers, and teachers. Finding eight people qualified to conduct observations is unrealistic for most schools.

Observer ratings did not appear to be a highly influencial factor on scores for Attending. Only 2% of the variance in scores was attributed to observer ratings. Additionally, interobserver agreement was slightly higher than chance to moderate in most observer pairs. Results indicated that scores for Attending were a reflection of student performance and not highly influenced by observer ratings.

Raise Hand

Raise Hand had a relative G-coefficient of .88; D-study results indicated that the reliability could be increased to as much as .93 with a combination of four occasions and six raters (this is an addition of one occasion and one rater to the present design). The additional resources required to obtain this level of reliability is not unreasonable. Progress-monitoring procedures support the monitoring of target skills twice a week; this would allow for at least four observations to be completed before making decisions regarding student progress on Raise Hand (e.g., after two weeks of observations). The number of observers required to achieve this level of reliability

may be higher than reasonable for a typical school. The level of reliability for the present design, however, was acceptable and no additional resources would need to be dedicated in order to achieve reliable results for this particular skill.

Observer ratings did not appear to be a highly influencial factor on scores for Raise Hand. Only 2% of the variance in scores was attributed to observer ratings. Additionally, interobserver agreement was slightly higher than chance to moderate in most observer pairs. Results indicated that scores for Raise Hand were a reflection of student performance and not highly influenced by observer ratings.

Hands to Self

A G-coefficient could not be calculated for Hands to Self because the data were homogenous. That is, all ratings for students on this skill were exactly the same, which resulted in a lack of variance for this skill. These results might indicate that the Hands to Self skill is not well-suited to this kind of behavioral observation. Student incidents of nonexamples of Hands to Self were fairly infrequent and might not be observed adequately through a momentary time sampling style of behavioral observation. Each observer rated each student with a perfect score for each occasion. This might indicate that observers were not sure what nonexamples of Hands to Self would resemble, and therefore, did not record them when they were present. Another explanation could be that students were less likely to engage in nonexamples of this behavior where most observations took place, in the classroom. It is also possible, as stated previously, that nonexamples are infrequent and not likely to be observed within a 15-minute period. Overall, results for Hands to Self cannot be interpreted in the same manner as the other observed skills.

Transition: Quiet

The relative G-coefficient for Transition: Quiet was .55. This is below an acceptable level of reliability. D-study results indicated that a combination of five occasions with eight observers would be expected to improve reliability to approximately .72 (an addition of two occasions and three observers to the present design). The addition of more than two occasions and three observers would be required to improve the reliability of Transition: Quiet to an acceptable level. As stated in the discussion of Attending, the addition of observers is likely to be more taxing on school resources than the addition of occasions. It is unlikely that a typical school would be able to find eight qualified observers to conduct behavioral observations twice a week for progress monitoring purposes.

Observer ratings accounted for nearly as much of the variance as did the object of measurement (students), which is likely why this particular skill received such low reliability results. Scores for Transition: Quiet may have been inconsistent, based on the low interobserver agreement (K = .161). Because training provided definitions of skills and practice observing skills, it may be that observers drifted away from protocol as time progressed or that training was not sufficient and observers never fully learned to identify the Transition: Quiet skill properly.

Transition: Follow Directions

The relative G-coefficient of.46 was below an acceptable level of reliability. D-study results indicated that reliability could be increased to .74 with a combination of five occasions and eightobservers (an increase of two occasions and three observers to the present design). As stated previously, the addition of raters would be taxing to

typical school resources. Observer accounted 0% of the variance for this skill, which indicates that observer ratings were not the largest factor influencing the reliability of this particular skill.

Observer agreement was slightly above chance for this skill (K = .25), which indicates that observers may have rated this skill based on their own definitions of examples and nonexamples. Much of the reliability was attributed to error in the residual term (93%), which could be explained, in part, by chance observer agreement.

Summary

Psychometric properties for the measure were mixed, indicating a good level of reliability for a behavioral measure (.80) and low percent of variance acccounted for by the object of measurement (11%). The Skill facet contributed to a larger than expected portion of the variance of the measure, and individual G and D studies were conducted for each skill respectively. Results from these studies indicated that a small portion of the variance for each skill was due to observer ratings. Overall, it appears that low interobserver agreement (or agreement close to chance levels) contributed to much of the variance in specific skill scores; this was an issue particularly for the Hands to Self and Transition: Quiet skills.

Cultural Considerations

Information on multiple demographic factors was collected for each participant, including age, gender, ethnicity, and SES. Individual analyses were conducted for each demographic factor to determine which skills might vary, and what between subjects differences could be found. Ethnicity stood out as the most prominent influencing factor on skills. Ethnicity had a large overall effect size of η^2

=.33. A followup ANOVA found a significant effect for the Transition-Quiet skill ($p < .05$) and a medium effect size of ($\eta^2 = .24$). Followup Tukey tests showed that Hispanic students were observed to perform more poorly on Transition-Quiet than White students. These findings are especially noteworthy due to the larger proportion of sample represented by Hispanic than White students. These findings have implications for how social skills are defined, observed, and measured in schools. It is important to understand the cultural impact that ethnicity may have on skill performance in order to reduce the overrepresentation of ethnic minorities in special education (Coutinho & Oswald, 2000). It is also important to interpret the present findings in a cautious manner; there is not enough information to determine ethnicity as an explanation for differences in skill performance. For example, the majority of students in the present sample were of Hispanic descent. It may be that Hispanic children were more talkative during transition activities because they were connected to a greater number of peers in their classroom than those from other ethnic backgrounds.

Implications

This study demonstrated the usefulness of employing G theory when developing a behavioral measure. This study showed how G theory can be used to validate a behavioral measure and assess the adequacy of specified measurement strategies. There are many behavioral-observation measures on the market that utilize technology such as iPhone applications but few, if any, have conducted studies in order to validate those measures. Validated instruments such as the Social Skills Improvement System (Gresham & Elliott, 2008) have gone through rigorous studies to

obtain reliability and validity for the measure, but do not utilize a technological framework that lends itself easily to progress monitoring.

The measure developed in this study, Metryx, could be used in a variety of ways. First, it could be used as an observational tool to collect progress-monitoring data on student performance of social skills, while saving and storing the collected data instantaneously in a secure database. Second, it could be used for decision making purposes to inform intervention instructors about student performance in certain skills, what skills have been mastered, and which require further or modified instruction. Third, it can be linked directly to a RTI format, informing teachers and instructors about where a student lies compared to personal goals/benchmarks (absolute comparison) and other students (relative comparison). Abilities this measure has to inform decision making in interventions as well as inform relative and absolute decisions are important for the area of social skils, where this kind of ability is lacking. This research adds to this area of study by establishing a reliable, efficient, and feasible measurement strategy to assess student performance of social skills. Comparable measures may be released to market without undergoing scrutiny to establish a reliable/valid measure as was done in the present study.

Although the measure's overall reliability was acceptable, the study indicated that changes should be made to the measurement strategy before use in progress monitoring. Baer, Harrison, Fradenburg, Petersen, and Milla (2005) reported that operational definitions of target behavior, time and setting of observations, and observational procedure (i.e., duration recording, momentary time sampling, partial-interval recording) should be carefully considered when using direct observation in

order to obtain reliable and valid results. Indeed, these factors would need to be addressed before using Metryx in schools for observing social skills. Specifically, the operational definitions of each skill should be clarified before observation, as they may have been unclear to observers in the present study, and time and setting of observations should be as consistent as possible, as they were not assessed as a facet in the present study, to reduce the influence they may have on behavior. The choice of observational procedure, momentary time sampling, in the present study would be appropriate for use in observing most skills as it is results in smaller estimation errors than partial or whole-interval observations (Hintze & Matthews, 2004). Consideration should be given to the choice of procedure and its appropriateness for the target skill for observation.

By establishing a reliable measure that is simple and efficient to use and easily lends itself to multiple decision making purposes, it is hoped that social skills instructors would use this measure to monitor the progress of student performance on skills and make appropriate adjustments in their instruction. By improving the appropriateness of instruction and differentiating it to student need, the effectiveness of social skills interventions may be improved and lead to higher student function in skill areas.

Limitations

Although G theory is less restrictive than CTT and considered multiple facets of the measurement design, there are still variables left unaccounted for. Data for 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. G theory also uses generalizability coefficients from one set of conditions (the G-study) and assumes that they apply to other predicted conditions (the D-study). This is a benefit of G theory, but also may be seen as a limitation when considering assumptions being made about similarity between conditions. It is important to remember that G theory derives predicted values and not obtained values and must be interpreted with some caution as a result.

A significant limitation to consider is the lack of reliability often found when using direct behavioral observation. Hintze and Matthews (2004) conducted a study in which better than adequate interobserver agreement was obtained as well as a high percentage of variance accounted for by the object of measurement, and still low levels of reliability were obtained (.60). They purport that this finding is not unique and that direct observation may not be as reliable of a method as the professionals who use it would like to believe.

The present study obtained an overall reliability coefficient of .80, an acceptable level of reliability for a behavioral measure. Salvia and Ysseldyke (2004) state that reliability coefficients of .90 or higher are recommended for instructional decision making purposes, but that coefficients of .70 or higher are recommended for screening purposes. Thus, the measurement strategy used in the present study was reliable enough for screening purposes (e.g., identifying students with similar needs for RTI purposes of instruction). However, as stated previously, the object of measurement accounted for only a small portion of the variance (11%). This indicates that the measurement strategy should be modified and not implemented as-is for progress-monitoring purposes. As stated previously, attention to operational

definitions, time and setting of observation, and observational procedure may improve the measurement strategy. It may be possible that matching an appropriate observational procedure to a specified skill, with clearly defined parameters, could decrease the amount of variance attributable to skills. Additionally, holding the time and setting of observation constant may reduce the amount of variance attributable to random error.

The reliability of an observed skill, or the likelihood that an observation accurately represents the true performance ability, improves with repeated observations across time (Hintze & Matthews, 2004). As was shown in many of the D-study results, a larger number of observations than was conducted in the present study would be necessary to obtain reliable results for specific skills. It may be that progress-monitoring of social skills may require more than the standard of semi-weekly observations for two weeks to accurately inform interventionists for decision making.

Observer ratings were an important factor in the present study. Interobserver agreement was variable; it varied from perfect agreement, slightly higher than chance, to moderate agreement. Due to a limited time frame for data collection, interobserver observations were conducted on one occasion only. A lack of data points made calculation of interobserver agreement for each skill difficult (and could not be calculated in many instances). Agreement may have been higher had there been more occasions of interobserver scores to evaluate. Another possibility for low interobserver agreement may have been due to insufficient training. Although each observer received three sessions of training and participated in multiple practice

observations prior to beginning the study, this training may not have been sufficient for the purposes of this study. Additionally, observers did not practice observing the same student in order to obtain reliability with each other before data collection began. This is one potential explanation for the observer drift that may have occurred. It is possible that this training was not enough and that more experienced observers with previous training in behavioral observation would have delivered more consistent ratings.

Another observer limitation that should be noted is the observation and interval length. because beginner observers were used, a longer interval was selected (i.e., 30 seconds) in order to obtain a more accurate score. The researcher determined to use a longer interval to reduce the effort needed to track interval length and hopefully obtain accurate scores for the appropriate interval. More experienced observers would be able to observe accurately using 15-second intervals while keeping track of time and student performance. Additionally, observations in the present study were 15-minutes in length, which might not have been enough time to obtain a representative sampling of student behavior on some skills. The researcher decided to have observers conduct 15-minute observations in order to maximize the number of students who could be observed during the limited time frame for data collection. This may be a limitation of the study design that could not be alleviated with the implementation of more experienced observers.

The observed skills are considered prerequisite social skills as they are fundamental for success in a classroom environment. These particular skills were chosen based on suggestion from school professionals at the participating school who

were working with students on some of the skills. These prerequisite skills were identified by teachers as important for children to master in the classroom. It is possible that other social skills may be more amenable to the type of observation used in the present study than the prerequisite skills that were observed. Additionally, each skill was evaluated differently (i.e., time sample or rubric). The rubric style of observation, for the prerequisite skills respectively, may have been more subjective and generated inflated scores. The method of observation, when using the measure is something that needs to be considered when assessing the best match between skill and measurement approach.

The setting in which observations took place is also worth consideration. Although time of day and type of activity in which the student was enged during observation were recorded, they were not held constant, or controlled for, or analyzed as a facet in the G-study. The time of day and type of activity occurring in the classroom could have had an impact on what skill performance looked like in the present study.

Finally, although the reliability of the measure used in this study was evaluated in detail for one particular population, generalizing to other populations should be done with caution. The researcher was fortunate to evaluate this measure with a population of children from primarily low-medium SES and ethnic minority backgrounds. This is a benefit for the study, but also means that information may not be appropriate to generalize to children from upper class ethnic majority backgrounds. Additionally, students in the present study attended a charter school, which means that generalizing to students attending public, private, or religious schools may not be

appropriate. It is possible that the observed population is qualitatively distinct from the student populations of other school settings.

Future Directions

The current study enhances the research in the area of social skills in several ways. By using G theory, a measurement strategy was developed that is reliable and, appropriately modified, can be used in schools effectively and efficiently. Results indicated that each skill may be qualitatively different from others and can be observed differently by trained individuals. Additionally, prerequisite skills may be qualitatively different than other social skills that may be commonly taught during interventions. Additional analyses could be conducted to determine the best measurement strategy for a number of social skills individually, using experienced observers. Research could also be done to account for environment and the impact it may have on skill performance for each of the observed skills. This would require a G-study for each skill, using environment, time of day, or day of the week as facets. Future research can use these findings to determine the best environment for teaching and measuring individual skills. Additionally, progress-monitoring procedures may need to be different from those frequently used in academic interventions in order to be reliable.

Social-skills interventions have been plagued by issues of ineffectiveness; students are often able to perform skills in the area of instruction but fail to generalize skills to other environments (Gresham, 2010; Gresham, Sugai, & Horner, 2001). Part of this problem may be due to the train and hope method of instruction where the instructor fails to adequately program for generalization and maintenance of skills and

simply hopes that the student will be able to utilize acquired skills in multiple outside settings (Gresham, Sugai, & Horner, 2001). The current study has evaluated a measure that could be used to increase effectiveness. The measure of interest, Metryx, allows an instructor to observe a skill, have immediate access to performance data, track student progress on multiple skills, compare performance to goal lines/benchmarks, and compare performance to other students. It is hoped that the use of this measure can improve effectiveness of social skills interventions by improving the decision making process for social skills instructors. Future research on this measure could increase the number of observers and enhance their training in order to obtain greater interobserver agreement as well as modify the grouping of assessed skills in order to facilitate a more homogenous conceptual understanding of social skill performance. Future research could use the measure in an experimental way to determine how this measure can be used during the decision making process and if that has an impact on intervention effectiveness. In addition, future research should investigate the use of this measure in multiple school settings with differing student populations in order to assess its appropriateness for various populations.

Summary and Conclusions

Social skills are an important aspect of student success in schools, both socially and academically (Cappadocia & Weiss, 2011; Cook, et al., 2008; Gresham, Cook, Crews, & Kern, 2004). Social-skills interventions have, historically, been reported to lack effectiveness as far as generalization of skills outside of the instructional environment (Gresham, 2010). One issue affecting the lack of effectiveness of social skills interventions is that many instructors fail to implement generalization instruction

into interventions (Gresham, Sugai, & Horner, 2001); little attention is given to tracking student progress or differentiating instruction. With the implementation of RTI in schools, more attention has been given to differentiating instruction to student need in order to improve student outcome and to monitor progress, with an emphasis on academic interventions, to gauge effectiveness of instruction (Stecker, Lembke, & Foegen, 2008). There is a need for a measure that can be used to monitor the progress of social-skill performance of students receiving services, one that can be implemented easily and efficiently and be linked directly to the RTI format. This study used G theory to study the reliability of such a measure.

The present study demonstrated the usefulness of G-theory for developing a multifaceted measurement strategy for behavioral observation. G-theory expands the traditional perspective of CTT by including multiple measurable facets to account for aspects of variance beyond random error. In addition, G-theory can be used to assess how different levels of each facet might affect the measure's reliability in hypothetical scenarios (D study). This study used G-theory to develop a measurement strategy for progress monitoring social skills, using prerequisite social skills for observation (attending, raise hand, hands to self, transition: quiet, transition: follow directions). The present study used student nested within observer as the object of measurement with skill, occasion, and observer as secondary facets.

The G study found that the measure demonstrated a good level of reliability for a behavioral measure with three occasions and five observers for obtaining a broad assortment of skills. Further analyses of each individual skill, however, revealed that a different measurement strategy would be beneficial depending on which skill was

being evaluated on an individual basis. D study analyses revealed that more than five occasions and eight observers may be required for some skills to be rated reliably. Other adjustments, such as modifying observational definitions, accounting for time and setting of observation, and ensuring observation be done by experienced observers, may need to be made in order to better account for student performance as the primary influence on obtained scores. Observer agreement was slightly higher than chance for some skills and may explain some of the low reliability. Reliability of certain skills may be improved by obtaining ratings from more experienced observers. It is possible that fewer observers may be required, as resources in schools are limited, if more experienced observers were used to provide more reliable ratings of skill performance.

This study developed a measure for monitoring progress of social skills performance that may be used in schools for decision-making purposes in order to improve the effectiveness of social skills interventions. Future research should seek to assess the reliability of this measure with different school settings and students from multiple backgrounds. Future research should assess how this measure can be used to inform the decision-making process and what impact its use may have on the effectiveness of interventions.

Appendix A:

Multicultural Considerations: A Brief Review of the Literature

It is important to consider socioeconomic, gender, and cultural factors that may have influenced the results of the current study. This appendix provides a review of the population from which the sample was drawn as well as a brief review of the research on socioeconomic, gender, and cultural factors related to social skills and classroom behavior. Awarenes of these factors is important in determining whether the method of evaluation is valid for diverse populations and whether the results of the study can be generalized.

It is reported that the majority of students at the participating school are Latino (45%) and African-American (31%). A small percentage of students at the participating school are White (17%), Asian (7%), and Native American (1%). About 3% of students at the participating school receive English language services and 13% receive additional educational supports through special education services. Approximately 60% of students at the participating school receive free or reduced lunch, which serves as an indicator of socioeconomic status (Infoworks, 2009).

Some research has indicated that African American students may receive less favorable treatment from their teachers compared to Caucasian students (Casteel, 1998). Less favorable treatment may be particularly detrimental to the social development of children from ethnically diverse backgrounds. For example, Burchinal, Peisner-Feinberg, Bryant, and Clifford (2010) found that the quality of care in childhood is differentially more important for children of color than for White non-Hispanic children. In the 2010 study, children from ethnically diverse backgrounds who received poor child care showed significantly impaired social behavior compared

to Caucasian students with similar experiences. Additionally, students from ethnically diverse backgrounds have historically been over-represented in special education (Coutinho & Oswald, 2000). These studies indicate the importance of examining cultural background when considering social-skills outcomes that may potentially differ by ethnicity.

Males are more often represented in social-skills groups than females, which may be due to a difference in the social development of males and females (Crombie, 1988). Taylor, Lian, Tracy, Williams, and Seigle (2002) found that teachers rated girls as more assertive than boys and rated males as lacking self-control more frequently than girls. This study indicates that similar behaviors may be interpreted differently based on teacher perception of appropriateness. This may be one reason that males are seen more often to be in need of social instruction. Crombie (1988) also suggests that children develop differently in responses to teacher and parent behavioral perceptions and expectations.

Socioeconomic status may be indirectly 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 percentage of children who qualify for free or reduced lunch is often used as an indicator of the SES of school populations). 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. These results indicate that children from low SES backgrounds may experience difficulties over time in academic and social areas if their basic dietary needs are not met.

Overall, the research indicates that there have historically been differences in the number of students referred for special education services from ethnically diverse backgrounds. Additionally, students from diverse backgrounds may be viewed by teachers to be lacking in social skills and treated less favorably as a result. Male students are often referred for social skills instruction more often than females. Lastly, students from low SES backgrounds may be at a disadvantage for developing healthy social skills due to situational factors. The population at the present school is composed of children from low SES, ethnically diverse backgrounds. Observational results should be valid and generalizable to schools with similar populations.

Appendix B: Informed Consent Forms

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 asking for permission to include your child in this study because they are students in the classrooms selected to participate in this study.

Description of the project:
Recently, your child's school has been working to develop a tool for teachers and other personnel at the school, called Metryx. Metryx has been used for tracking student progress in the classroom and keeping track of academic files. The purpose of this project is to see if Metryx can be used as well for observing how young children learn typical classroom skills such as raising their hand before being called on and following directions.

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 the children 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.

Risks or discomfort:
There are no risks or discomfort involved for your child in this project. It will be explained to the class that there will sometimes be a person from URI observing the classroom so that the children 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 Metryx as well as the services they will be able to provide to students.

Confidentiality:
Your child's part in this study is confidential. All information will be stored electronically in the Metryx system, 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:
Children 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 Name Charter School. 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 Professor W. Grant Willis (401) 874-4245 or with Graduate Student, Monica Mabe (435) 760-7213, both from URI. Key personnel from Name Charter School involved in this study are School Psychologist Dehlia McCarthy (401) 277-2600, Occupational Therapist Tania Rosa (401) 277-2600, and Metryx CEO Shawn Rubin (401) 831-7323. 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.

_____	_____
Signature of Participant	Signature of Researcher
_____	_____
Typed/printed Name	Typed/printed name
_____	_____
Date	Date

Appendix C: Student Assent Form

Student Assent Form
(to be read aloud to potential participants)

Some college students from University will be coming to your school to learn about children. Your teacher and your parents already know about this. Is it OK if the URI students watch you for a little while when they are here?

_____ Yes

_____ No

My name is _____

Bibliography

Ang, R. P., & Hughes, J. N. (2001). Differential benefits of skills training with antisocial youth based on group composition: A meta-analytic investigation. *School Psychology Review*, 31, 164-185.

Baer, D. M., Harrison, R., Fradenburg, L., Petersen, D., Milla, S. (2005). Some pragmatics in the valid and reliable recording of directly observed behavior. *Research on Social Work Practice*, 15, 440-451.

Beelmann, A., Pfingsten, U., & Losel, F. (1994). Effects of training social competence in children: A meta-analysis of recent evaluation studies. *Journal of Clinical Child Psychology*, 23, 260-271.

Burchinal, M. R., Peisner-Feinberg, E., Bryant, D. M., & Clifford, R. (2010). Children's social and cognitive development and child-care quality: Testing for differential associations related to poverty, gender, or ethnicity. *Applied Developmental Science*, 4, 149-165.

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.

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.

Casteel, C. A. (1998). Teacher-student interactions and race in integrated classrooms. *The Journal of Educational Research*, 92, 115-120.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.

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.

Coutinho, M. J., & Oswald, D. P. (2000). Disproportionate representation in special education: A synthesis and recommendations. *Journal of Child and Familty Studies*, 9, 135-156.

Crombie, G. (1988). Gender differences: Implications for social skills assessment and training. *Journal of Clinical Child Psychology*, 17, 116-120.

Durlak, J. A., Fuhrman, T., & Lampman, C. (1991). Effectiveness of cognitive-behavior therapy for maladapting children: A meta-analysis. *Psychological Bulletin*, 2, 204-214.

Fletcher, J. M., & Vaughn, S. (2009). Response to Intervention: Preventing and remediating academic difficulties. *Child Development Perspectives*, 3, 30-37.

Fuchs, L. S., & Fuchs, D. (2009). On the importance of a unified model of responsiveness to intervention. *Child Development Perspectives*, 3, 41-43.

Furr, R. M. & Bacharach, V. R. (2008). *Psychometrics: An introduction.* Thousand Oaks, CA: Sage Publications.

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.

Hintze, J. M., Christ, T. J., & Methe, S. A. (2006). Curriculum-based assessment. *Psychology in the Schools*, 43, 45-56.

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.

Information Works (2009). Using Information, Name Charter School. Retrieved April 16, 2012 from http://www.infoworks.ride.edu/2009/pdf/usinginfo/28601E-info.pdf.

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.

Losel, F., & Beelman, A. (2003). Effects of child skills training in preventing antisocial behavior: A systematic review of randomized evaluations. *Annals, AAPSS*, 857, 84-109.

Mushquash, C. & O'Connor, B. P. (2006). SPSS, SAS, and MATLAB programs for generalizability theory analyses. *Behavior Research Methods*, 38, 542-547.

Quinn, M. M., Kavale, K. A., Mathur, S. R., Rutherford, R. B., & Forness, S. R. (1999). A meta-analysis of social skill interventions for students with emotional or behavioral disorders. *Journal of Emotional and Behavioral Disorders*, 7, 54-64.

Salvia, J., & Ysseldyke, J. E. (2004). *Assessment* (9th ed.). Princeton, NJ: Houghton Mifflin.

Schneider, B. H. (1992). Didactic methods for enhancing children's peer relations: A quantitative review. *Clinical Psychology Review*, 12, 363-382.

Schneider, B. H., & Bryne, B. M. (1985). Children's social skills training: A meta-analysis. In B. H. Schneider, K. H. Rubin, & J. E. Ledingham (Eds.), *Children's peer relations: Issues in assessment and intervention* (pp. 175-192). New York, NY: Springer-Verlag.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer*. Thousand Oaks, CA: SAGE Publications.

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.

Stemler, Steven E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4). Available online: http://pareonline.net/getvn.asp?v=9&n=4.

Taylor, C. A., Liang, B., Tracy, A. J., Williams, L. M., & Seigle, P. (2002). Gender differences in middle school adjustment, physical fighting, and social skills: Evaluation of a social competency program. *The Journal of Primary Prevention*, 23, 259-272.

Villares, E., Brigman, G., & Peluso, P. R. (2008). Ready to learn: An evidence-based individual psychology linked curriculum for prekindergarten through first grade. *The Journal of Individual Psychology*, 64, 403-415.