

2013

# Generalizability and dependability of brief behavior rating scales for social skills

Lisa L. Minor

Louisiana State University and Agricultural and Mechanical College, llibster@gmail.com

Follow this and additional works at: [https://digitalcommons.lsu.edu/gradschool\\_dissertations](https://digitalcommons.lsu.edu/gradschool_dissertations)

 Part of the [Psychology Commons](#)

## Recommended Citation

Minor, Lisa L., "Generalizability and dependability of brief behavior rating scales for social skills" (2013). *LSU Doctoral Dissertations*. 281.

[https://digitalcommons.lsu.edu/gradschool\\_dissertations/281](https://digitalcommons.lsu.edu/gradschool_dissertations/281)

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Doctoral Dissertations by an authorized graduate school editor of LSU Digital Commons. For more information, please contact [gradetd@lsu.edu](mailto:gradetd@lsu.edu).

GENERALIZABILITY AND DEPENDABILITY OF BRIEF BEHAVIOR  
RATING SCALES FOR SOCIAL SKILLS

A Dissertation

Submitted to the Graduate Faculty of the  
Louisiana State University and  
Agricultural and Mechanical College  
in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

in

The Department of Psychology

by

Lisa Libster Minor  
B.A., Towson University, 2004  
M.A., Louisiana State University, 2009  
December 2013

## ACKNOWLEDGMENTS

To Dr. Frank M. Gresham, for his unwavering support over the past seven years. It has truly been inspirational working with him. I have yet to write a paper without discovering that he has already written a paper or grant on the topic. I am forever grateful to Dr. Gresham (and Laura Gresham) for demonstrating that a work-life balance can actually be achieved. Indeed, I now believe tailgating and crawfish boils to be mandatory experiences when attending graduate school in South Louisiana. Geaux Tigers!

I am similarly indebted to Dr. Lynn Singletary, who has served as a role model and mentor during my graduate school career. I could not imagine coming this far without her support, gentle guidance, and tough love when necessary.

I also wish to thank members of my committee: Dr. Noell, Dr. Davis, and Dr. Van Gemmert for their support with this project. I am also beholden to Dr. R. J. Volpe of Northeastern University for serving as a subject matter expert on generalizability theory applications to social behavior assessments and as a sounding board during project development.

I am grateful to my fellow former school psychology graduate students for support, especially Dr. Michael Vance and Dr. Keri Menesses. I could not have completed this journey without them.

Finally, I wish to thank my friends and family for their support. To my parents, Stephen and Enid Libster, for providing every sort of support possible for me to achieve my goals. To my best friends, Stacey Leonard for her shared love of books, education and behaviorism, and to Dr. Christina Ragan for blazing the PhD trail ahead of me.

Last, but certainly not least, to my husband, Dr. Kyle Minor, whom I met at the psychology graduate student mixer before our first week of classes. I have now gone through the

entirety of my graduate career and burgeoning professional career with him by my side. Thank you for inspiring me to achieve my best and for modeling dedication and passion for work.

## TABLE OF CONTENTS

Acknowledgments.....	ii
List of Tables.....	vi
List of Figures.....	vii
Abstract.....	viii
Introduction.....	1
Generalizability Theory.....	5
Relevance of Classical Test Theory to Behavioral Assessment.....	5
Generalizability Theory.....	8
Generalizability Studies.....	9
Decision Studies.....	11
Reliability and Validity in G Theory.....	12
Statistics.....	13
Relevance of G Theory to Behavioral Assessment.....	20
Review of Methods Used To Assess Social Behavior.....	25
Systematic Direct Observation.....	25
Behavior Rating Scales.....	29
Direct Behavior Ratings.....	32
Brief Behavior Rating Scales.....	40
Purpose and Rationale.....	48
Method.....	53
Participants and Setting.....	53
Measures.....	54
Procedures.....	57
Statistical Procedures and Analyses.....	61
Generalizability Studies.....	61
Decision Studies.....	63
Results.....	70
Generalizability Theory Analyses.....	70
Discussion.....	84
Summary of Statistical Findings.....	84
Implications of Findings.....	91
Relevance of Current Research to Previous Literature.....	95

Recommendations to Address Rater Biases.....	98
Recommendations for Future Scale Development.....	102
Limitations.....	104
Conclusion.....	107
References.....	108
Appendix: Institutional Review Board Approval Sheet.....	120
Vita.....	121

## LIST OF TABLES

1.	Sources of Variation in $p \times i \times o$ Fully Crossed Random-Model Design.....	15
2.	Reliability Estimates as Reported in the Social Skills Improvement System Rating Scale Manual for the Teacher Form.....	56
3.	Sources of Variation in $p \times r \times I \times X \times o$ Fully Crossed Mixed-Model Design with Items Fixed.....	62
4.	G-Study Variance Components Using ANOVA with Type III Sums of Squares.....	64
5.	Variance Notation for G and D Studies, Averaging Across Fixed Facet (Items) Method.....	66
6.	Descriptive Statistics For Items Across Occasions.....	71
7.	Full Model G Study Variance Component Estimates.....	72
8.	Reduced Model G Study Variance Component Estimates for Communication.....	78
9.	Reduced Model G Study Variance Component Estimates for Cooperation.....	79

## LIST OF FIGURES

1.	Venn Diagrams Displaying Sources Of Variance For p X R X I X O Fully Crossed Random Effects Model Study Design.....	21
2.	Venn Diagrams Displaying Sources of Variance For p X R X I X O Fully Crossed Mixed Model Study Design, with I Fixed and R and O Random.....	22
3.	Participant Selection Flowchart.....	59
4.	Mean Frequency Ratings by Rater for the Full 7-Item Communication and 6-Item Cooperation Subscales.....	75
5.	Full Model Dependability Studies.....	76
6.	Generalizability and Dependability Coefficients for the Communication Subscale. ....	82
7.	Generalizability and Dependability Coefficients for the Cooperation Subscale.....	83



## ABSTRACT

While there are appropriate tools to progress monitor academics, there is no universally accepted tool to progress monitor social behavior. The current study proposed the development of a series of brief behavior rating scales to correspond to important social skills domains on the Social Skills Rating Scale (Gresham & Elliott, 2008) and to evaluate the resulting psychometric features through generalizability theory. Data was collected in a preschool classroom in a 6 persons by 2 rater by 6-7 items by occasions mixed model design. Data was analyzed series of generalizability and decision studies to investigate sources of variability and to determine the assessment conditions required to make reliable decisions. Results indicated that large proportions of the total variance in these scales were attributed to rater-related effects. This affected generalizability and dependability coefficients to the extent that reliable decisions could not be made with the current scale using feasible assessment conditions. Furthermore, current results did not support the abbreviation of the current scales. These finding indicates the need to further understand and control for unwanted variability between raters. Implications for the assessment of social behavior and suggestions for scale development are reviewed.

## INTRODUCTION

In recent years, school psychology has undergone a paradigm shift from the traditional wait-to-fail model for academic and behavioral problems to problem-solving models. These include response to intervention (RTI) and positive behavior supports (Gresham, 2004; Tilly, 2008). The traditional model holds that students must demonstrate severe discrepancies between expected and actual performance before qualifying for services. Problem-solving models focus on prevention and the provision of high quality interventions across multiple levels of supports (Basche et al., 2005; Gresham, 2002a). RTI is primarily a service delivery model in which the intensity of the intervention is matched to the intensity of the academic or behavioral problem (Gresham, 2004; Jimerson, Burns, & VanDerHeyden, 2007).

A core feature of RTI is its reliance on data-based decisions to determine if students are adequately responding to intervention efforts (Gresham, 2008). RTI requires appropriate assessment tools that can identify students who fail to make adequate progress, establish rate of improvement, and inform decisions such as whether to modify, terminate, or continue the current intervention (Gresham et al., 2010). As noted by Volpe and Briesch (2012), schools are increasingly adopting RTI for academics, in part due to the wide base of research supporting curriculum-based measurement (CBM) as a feasible assessment tool. In comparison, adoption of RTI for social behavior has been much slower, perhaps due to the lack of a CBM analogue for evaluating social behavior. This is unfortunate, as social behavior is very relevant to school settings. Social behavior consists two related domains: social skills and problem behavior (Gresham et al.). Numerous studies have linked social skills and problem behavior to academic achievement (Caprara, Barbaranelli, Pastorelli, Bandura, & Zimbardo, 2000; DiPerma & Elliott, 2002; Malecki & Elliott, 2002; Wentzel, 1993).

Furthermore, failure to develop social competency has been linked to a host of immediate and long term negative outcomes across psychological, social and educational domains (Kupersmidt, Coie, & Dodge, 1990; Newcomb, Bukowski, & Patee, 1993; Parker & Asher, 1987). Accordingly, difficulties in areas related to social competency and adjustment are a diagnostic feature of several of DSM-IV disorders and disability classifications under the Individuals with Disabilities Improvement Education Act (American Psychiatric Association, 1994; IDEA, 2004). Of further concern, while 22% of students require mental health interventions (Hoagwood & Erwin, 1997), the vast majority of these students do not receive services (Gresham, 2004). While many interactive and additive risk factors contribute to these outcomes, factors related to social competencies represent malleable targets for intervention that can be addressed within school settings (Crews, Bender, Gresham, Kern, Vanderwood & Cook, 2007).

Given the importance of social behavior within school settings, it is clear that preventive models are needed. As noted by Volpe and Gadow (2010), the success of these models depends on the establishment of reliable and feasible behavioral assessment methods. Chafouleas, Volpe, Gresham and Cook (2010) describe four requirements of behavioral assessment methods to problem-solving contexts: defensibility, flexibility, efficiency and repeatability. Defensibility refers to documented evidence for the reliability and validity of a measure. Flexibility is demonstrated when the measure can be used in a variety of situations or different purposes. Efficiency refers to the resources and time required to complete the assessment; in other words, the feasibility of assessment procedures. Finally, the measure must be repeatable, meaning that it can be administered in a formative manner to assess response to intervention.

Applied to academic problem-solving models, CBM clearly meets these objectives (Shinn, 2002). CBM is an assessment tool, which can be used to identify students at risk for academic failure (e.g., students who fail to make adequate progress towards established benchmarks), to monitor progress in response to evidenced-based instruction, and to facilitate instructional decisions such as whether additional intervention is warranted (Shinn).

When applied to social behavior, there is no tool that clearly accomplishes all of these objectives. While excellent methods exist for screening and identifying social skills deficits and problem behaviors, such as the Systematic Screening for Behavior Disorders (Walker & Severson, 1990) and the Social Skill Improvement System Rating Scales (Gresham & Elliott, 2008), progress monitoring tools are still being developed (Chafouleas et al., 2010). As aptly stated by Gresham (et al., 2010, p.365), “There is no ‘CBM analogue’ for dependably measuring students’ response to short-term interventions in the area of social skills and problem behaviors.” While this is certainly true, drawing parallel with CBM may be useful in that CBM is not a single tool, but rather a methodology consisting of three types of tools: general outcome measures (GOM), skill-based measures, and mastery measures. These various types of CBM serve different purposes. For example, GOM is a dynamic indicator of overall functioning and is used for screening and progress monitoring purposes, whereas skill-based measures and mastery measures can be used to identify or progress monitor specific skill areas (Hosp, Hosp & Howell, 2007). Following Kratochwill and Bergan’s (1990) recommendations, Volpe and colleagues suggest that socio-behavioral performance should be monitored on both broad and specific levels (Briesch & Volpe, 2007; Volpe & Briesch, 2012; Volpe & Gadow, 2010).

In recent years, four behavioral assessment methods have been commonly used to assess social behavior: systematic direct observation (SDO), behavior rating scales (BRS), direct

behavior ratings (DBR) and brief behavior rating scales (BBRS). Each method will be reviewed in detail in the third chapter of this manuscript. In the past, social behavior was predominately assessed with SDO and BRS. As will be discussed, these methods may not be tenable for frequent progress monitoring purposes as required by RTI (Briesch & Volpe, 2007; Christ, Riley-Tillman, & Chafouleas, 2009). DBR and BBRS have been recently proposed as progress-monitor tools for social behavior. However, these methods are still in the early stages of development and further research is required to substantiate the psychometric properties and feasibility of these methods.

Drawing on lessons learned from developing CBMs, Fuchs (2004) suggested that progress-monitoring tools be developed in three stages. The first stage consists of evaluation of the “*technical features of the static score.*” This refers to formal evaluation of the reliability and validity of the measure, assessed at one point in time. The second stage is evaluation of the “*technical features of the slope.*” At this stage, the measure should demonstrate change-sensitivity; in other words, the measure should be able to capture behavioral changes that occur in response to the intervention. Finally, in the third stage, the measure must demonstrate “*instructional utility*” (Fuchs, p. 189). At this point, the measure should demonstrate the ability to facilitate data-based decisions such as whether to modify or titrate the intervention. Therefore use of this data should produce improved outcomes for students.

In the current study, we propose to develop a series of brief behavior rating scales to assess social behavior. As a new measure (and indeed, one of only a handful of BBRS measures designed specifically for use as progress monitoring tool within an RTI context), the focus of this study is on the first stage of development. This will be accomplished through generalizability theory.

## GENERALIZABILITY THEORY

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999) and well as the Individuals With Disabilities Education Improvement Act (2004) require the use of tests with sound psychometric features. Classical Test Theory (CTT) has long been used to establish the reliability of educational and psychological measures. The psychometric properties of intelligence and achievement tests have been thoroughly evaluated while the psychometric properties of behavioral assessment measures have received far less attention (Briesch, Chafouleas, & Riley-Tillman, 2010). As noted by Hintze (2005), this may be due to the perceived limited relevance of CTT to traditional forms of behavioral assessment (i.e., direct observation). With the recent interest in developing tools to assess social behavior, alternative methods, especially generalizability theory are increasingly used to establish the reliability of behavioral assessment measures. To provide context for the current study, this section will review the limitations of CTT to behavioral assessment and then outline generalizability theory (G theory) as an important alternative for establishing the psychometric features of behavioral assessment methods.

### **Relevance of Classical Test Theory to Behavioral Assessment**

Classical Test Theory and behavioral assessment operate under different assumptions that are often difficult to reconcile. A primary assumption of CTT is that individuals possess stable traits or characteristics (Ghiselli, Campbell & Zedeck, 1981). Under this approach, an observed score is considered as a sign of the underlying latent trait, as such behavior should be stable across time and settings (Gresham & Carey, 1988). In contrast, the key assumption of the behavioral assessment approach is that the individual's behavior is a product of his/her learning

history. In other words, behavior is environmentally determined, situationally specific and malleable (Kazdin, 1979; Nelson, 1983). A behavioral observation is considered to be a sample of behavior that occurs in similar situations (Godfried & Kent, 1972).

In CTT, data are used to diagnose, classify or potentially to evaluate treatment in a pre-post fashion (Hartmann, Roper, & Bradford, 1979). In contrast, behavioral assessment focuses on the repeatability of measurement, as data are used in a formative matter throughout treatment. Furthermore, traditional assessment only considers relative comparisons between individuals, whereas behavioral assessment considers both within and between individual comparisons. Finally, CTT uses high inference measures, reflecting a focus on latent traits such as intelligence and aptitudes whereas behavioral assessment relies on low-inference measures to assess observable behaviors (Godfried & Kent, 1972; Traub, 2005).

While acknowledging these arguments, Gresham and Carey (1988) attempted some reconciliation between the two models. They noted that when behavior is modified in a controlled fashion, it should produce systematic changes in behavior. In CTT systematic variation is considered as part of the true score rather than random measurement error. As such, this does not violate the assumptions of CTT. Despite these benefits, the assumptions of CTT remain ill-suited towards establishing the psychometric properties of progress monitoring tools. Again, the goal of a progress-monitoring tool is to evaluate responsiveness to the application of an intervention. Thus, a progress monitoring tool should be able to account for systematic variations in behavior over time and setting (Cone, 1977; Volpe & Chafouleas, 2011).

Another major limitation of CTT is that a measure may possess multiple and contradictory reliability coefficients when different procedures are used to calculate reliability

(Webb, Shavelson, & Haertel, 2006). As observed by Goodenough (1936; as cited in Cronbach, Gleser, Nanda, & Rajaratnam, 1972):

The investigator who compares two administrations of the same list of spelling words asks a different question than the investigator who compares performance on two different lists. Inconsistency of observers, inconsistency in the subject's response to different stimulus lists, inconsistency of his response to the same stimulus on different occasions- all of these may be sources of error, but any one comparison will detect some inconsistencies and not others.

Although it is common practice to use different methods to calculate reliability to account for these various sources of error, Suen and Pui-Wa (2007) described several conceptual and statistical problems that arise with this practice. Conceptually, CTT can only accommodate one source of random error at a time and assumes that all other sources of variance are fixed. As noted by Brennan (2011, p. 7), “[this] does not mean that there is necessarily only one source of error, however, within a single application of CTT, all sources of error are confounded in one *E* term.” Statistically, different methods of calculating reliability result in different standard errors of measurement (SEMs). Therefore, it is unclear which SEM should be used to construct confidence intervals around observed scores (Brennan). Theoretically, averaging across the various SEMs could solve this issue, but CTT offers no mechanism to combine various sources of measurement error (Suen & Pui-Wa).

A further limitation of CTT is that obtained psychometric values are highly dependent on the test and sample (Christ & Hintze, 2007). This relates back to the parallel test assumption, which states that two tests are parallel when the same true score and error variances are obtained across equivalent forms of the test (e.g. across items, testing conditions, administrations). Changing aspects of the measurement situation may result in different true scores and obtained values. Thus, a criticism of CTT is that psychometric values only apply to the narrow measurement conditions under which the test was originally administered, including the



characteristics of the norming sample (Chafouleas, Christ, Riley-Tillman, Briesch and Chanese, 2007). For this reason, test developers must carefully select the standardization sample to be representative of the population they intend to measure. Moreover, test users must ensure that the test has been validated for the population they intend to test (AERA, APA, & NCME, 1999).

### **Generalizability Theory**

Due to the limited relevance of CTT to behavioral assessment, generalizability theory (G theory) has been advocated as an alternative approach to establish the technical adequacy of assessment of social behavior (Cone, 1977; Gresham & Carey, 1988; Hintze, 2005). The foundations of G theory were developed by Cronbach, Glaser, Nanda, and Rajaratnam (1972) in their seminal monograph, *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Of note, the original intent of this work was to develop a CTT handbook. G theory arose when Cronbach was unable to reconcile the contradictions to reliability as described in the preceding section (Cronbach, 1991 as cited in Brennan, 2011).

G theory is an extension of CTT that can account for multiple sources of variance simultaneously (Brennan, 2010; 2011). Whereas in CTT measurement error is assumed to be random, in G theory, additional sources of systematic variance can be accounted for using repeated-measures factorial ANOVA. This allows for calculation of variance associated with different conditions, or facets of measurement, such as rater, setting, method, occasion, etc. These variance estimates are used to investigate the reliability of outcomes under various measurement scenarios. This information can be used to produce more efficient measurement procedures and thus improve measurement decisions (Webb, Shavelson, & Haertel, 2006). The Standards for Educational and Psychological Testing (American Educational Research, 2002)

recommend the use of G theory for determining error variances that arise from multiple sources, when feasible.

Shavelson and Webb define G theory as, “the statistical theory about the dependability of behavioral measurements” (1991, p. 1). Dependability refers to the accuracy with which one can generalize from a person’s observed score on a measure to the average score obtained across all possible acceptable measurement conditions (e.g., similar behaviors, raters, settings, methods, etc.).

### **Generalizability Studies**

The purpose of a generalizability study (G study) is to identify the variance associated with the object of measurement and with the various conditions, or facets of measurement, and interactions between these facets. Each facet represents a potential universe of similar measurement conditions. Examples include similar raters, settings, methods, occasions, or items (Brennan, 2010; Shavelson & Webb, 1991). A universe is described as all possible acceptable instances of the condition of measurement (e.g., all similar test items, or all similar assessment occasions). The variance associated with the object of measurement (often *person*) reflects true differences between on the measured trait or behavior (similar to the true score in CTT). The variance due to facets and interactions reflects error variance associated with the conditions of measurement (Brennan, 2011).

An important task within a G study is the definition of the universe of admissible observations. This consists of all observations across the various conditions or facets of measurement that the user is willing to define as interchangeable (Shavelson & Webb, 1991). Often, the researcher is interested in multiple facets such as generalizing from the observations of one teacher to other teachers and generalizing from the observed items to the larger pool of

items within a domain of interest. Shavelson and Webb recommend that the universe of admissible observations be defined as broadly as possible to facilitate multiple D studies from the data.

Once defined, the researcher collects data using the specified conditions as a representative sample of all possible observations within the universe of admissible observations. These data are analyzed using repeated measures factorial ANOVA. This allows the observed scores to be decomposed and used to derive variance estimates associated with each component (Brennan, 2010; Shavelson & Webb, 1991). Unlike traditional uses, ANOVA is used solely to estimate the variance components associated with each facet (Brennan).

As noted by Shavelson and Webb, the variance components for a given facet reflect the main effect of the condition of measurement (e.g., constant effect for all persons due differences in item difficulty), whereas the variance components for interaction terms reflect inconsistencies across different levels of facets (e.g., some items are more difficult for certain students than others). Once obtained, estimated variance components are used in D studies to estimate the universe score and error variances, which are used to calculate reliability like- coefficients (Brennan, 2011).

**G and D study designs.** Before continuing, it is helpful to review some G theory terms. *Facet* refers to a set of similar measurement conditions, e.g., similar raters, or similar test forms. Facets may be specified as random or fixed, depending on the generalizations that user wishes to make. When specifying a random facet, the user considers the current levels of the facet as a random sample of that universe. In specifying a fixed facet, the user will not generalize beyond the observed levels of the facet. Statistically, the fixed facet does not independently contribute to error variances (Cardinet, Johnson, & Pini, 2009). A random effect design indicates that all

facets are random, and a mixed model design means that some facets are fixed (Brennan, 2010; Shavelson & Webb, 1991).

The G or D study design describes the number of facets [note that object of measurement, typically person, is not included when describing the number of facets], and whether these facets are crossed or nested. A design is considered fully crossed when every condition of each facet is encountered by every condition of all other facets. In nested designs, conditions of a given facet are only experienced by some conditions of other facets. When possible, it is recommended to use a fully crossed design, as the nested facets are confounded (i.e., can not be interpreted separately), which may reduce the total explained variance relative to a similar fully crossed design (Chafouleas, Briesch, Riley-Tillman, Christ, Black, & Kilgus, 2010).

### **Decision Studies**

In decision studies (D studies), the estimated variance components obtained in the G study are used to design more efficient measurement procedures for a specified purpose (Brennan, 2010). For example, in a G study, a large error variance component associated with items suggests that the observed items varied in difficulty. In the D study, a researcher can evaluate if adding more items will reduce the error variance and thus increase the sensitivity to measuring true individual differences (Hintze, 2005).

Similar to reliability in CTT, the degree to which the assessment scenario is able to capture true differences (and thus improve decisions) is quantified as a generalizability or dependability coefficient (Shavelson & Webb, 1991; Suen & Pui-Wa, 2007). Whereas CTT can only be used to make relative decisions based on the relative ranking of individuals, G theory can be used to make both relative and absolute decisions about the individual's absolute level of performance, reflecting the generalizability and dependability coefficients respectively

(Shavelson & Webb). D studies allow the user to investigate how changes to measurement situation (can be used to improve the decisions based on measurement outcomes (Hintze & Mathews, 2004). D studies can also be used to determine the measurement conditions required to obtain adequate psychometric features. Within the recent behavioral assessment literature, G theory has been used to determine how many rating occasions or observations are needed to obtain dependable decisions (Briesch et al., 2010; Chafouleas et al., 2007; Christ, Riley-Tillman, Chafouleas & Boice, 2010; Hintze & Mathews, 2004; Volpe et al., 2011).

An important task in conducting a D study is the specification of the universe of generalization. The universe of generalization refers to the population of all possible measurement conditions that are of interest when making decisions for a particular purpose (Crocker & Algina, 1986). Furthermore, multiple universes of generalization can be specified for different purposes using the same G study variance component estimates (Brennan, 2003; Brennan, 2010). The sample sizes for each facet are not required to be the same as those used in the G study.

### **Reliability and Validity in G Theory**

G theory retains many parallels to reliability and validity within CTT while blurring the distinctions between these concepts (Brennan, 2000; Cronbach et al., 1972). The universe score variance in G theory is conceptually similar to the true score in CTT. The generalizability coefficient, expressed as the ratio of universe score variance to observed score variance (universe score variance plus relative error variance) directly parallels the reliability coefficient in CTT. The dependability coefficient replaces relative error variance with absolute error variance for absolute decisions (Brennan).

Cone (1977) defined six universes of generalization that are applicable to behavioral assessment: scorer, item, time, setting, method, and dimension. Many of these have parallels to other conceptualizations of reliability and validity. For example, scorer generalizability is similar to inter-observer reliability or inter-observer agreement. Along these lines, time generalizability mirrors test-retest reliability. Setting generalizability is like criterion-related validity, items generalizability is similar to internal consistency or even construct validity, method generalizability approximates convergent validity, and dimension generalizability can represent discriminate validity (Cone; Gresham & Carey, 1998).

A key distinction between the conceptualization of reliability in G theory and CTT is that the former considers reliability and error within the context of the measurement situation (Hintze & Mathews, 2004). In G theory, variance components are estimated directly from the observed data. These components are used as “building blocks” from which separate reliability-like coefficients can be estimated for different measurement scenarios (Suen & Pui-Wa, 2007, p.4). Examples include using different raters, or increasing or decreasing the number of observations. Thus, unlike CTT, the obtained reliability and error variances are not test or sample dependent.

## **Statistics**

The following section provides a basic overview of the statistical underpinnings of G theory. For more advanced understanding, Generalizability Theory (Brennan, 2010), Generalizability Theory: A Primer (Shavelson & Webb, 1991) are excellent resources. For illustration, we will describe  $p \times i \times o$  random effects model with a fully crossed design and discuss how this differs from similar design with  $i$  as a fixed facet.

**G study statistics.** The statistical task of a G study is to decompose a person’s observed score into its component parts to estimate the variance components. This is accomplished

through using repeated measures factorial ANOVA, with person, item, and occasion as main effects, person x item, person x occasion, and item x occasion as interaction terms, and person x item x occasion as the residual term. This can be expressed in a linear model:

$$X_{pio} = \mu + v_p + v_i + v_o + v_{pi} + v_{po} + v_{io} + v_{pio} \quad (1)$$

where  $\mu$  is the grand mean and  $v$  is the effect of the component or:

$$\begin{aligned} X_{pio} = & \mu && \text{[grand mean]} \\ & + \mu_p - \mu && \text{[person effect]} \\ & + \mu_i - \mu && \text{[item effect]} \\ & + \mu_o - \mu && \text{[occasion effect]} \\ & + \mu_{pi} - \mu_p - \mu_i + \mu && \text{[p x i interaction effect]} \\ & + \mu_{po} - \mu_p - \mu_o + \mu && \text{[p x o interaction effect]} \\ & + \mu_{io} - \mu_i - \mu_o + \mu && \text{[i x o interaction effect]} \\ & + \mu_{pio} - \mu_{pi} - \mu_{po} - \mu_{io} + \mu_p + \mu_i + \mu_o - \mu && \text{[residual effect]} \end{aligned} \quad (2)$$

where  $\mu$  is the grand mean across population and universe of admissible observations,  $\mu_p$  is the universe score, or the person's average score over the entire universe of admissible observations,  $\mu_i$  is the population's average score for the given item, and  $\mu_o$  is the population's average score for the given occasion (Brennan, 2010; Shavelson & Webb, 1991).

The total variance across all people and facets within the universe of admissible observations can also be expressed in a linear fashion as:

$$\sigma^2(X_{pio}) = \sigma_p^2 + \sigma_i^2 + \sigma_o^2 + \sigma_{pi}^2 + \sigma_{po}^2 + \sigma_{io}^2 + \sigma_{pio}^2 \quad (3)$$

(Brennan, 2010). Table 1 lists the seven components, or sources of variance in this design.

The estimated variance for each component is calculated using factorial repeated-measures ANOVA. The degrees of freedom are equal to the product of  $(n_\alpha - 1)$  for all indices

Table 1  
Sources of Variation in  $p \times i \times o$  Fully Crossed Random-Model Design

Facet	Type of Variation	Variance Notation
Persons	Object of measurement	$\sigma_p^2$
Items ( $i$ )	Variability due to inconsistency in item difficulty	$\sigma_i^2$
Occasions ( $o$ )	Variability due to inconsistency across occasions	$\sigma_o^2$
$p \times i$	Variability due to $p \times i$ interaction	$\sigma_{pi}^2$
$p \times o$	Variability due to $p \times o$ interaction	$\sigma_{po}^2$
$i \times o$	Variability due to $i \times o$ interaction	$\sigma_{io}^2$
$p \times i \times o + e$	Variability due to $p \times i \times o$ interaction plus error	$\sigma_{pio}^2$

that make up the given effect, with  $n_\alpha$  representing the G study sample size for that facet (Brennan, 2010). For example, for a  $p \times i \times o$  study with 10 participants, 5 items and 3 occasions, the degrees of freedom for the  $p$  effect is 9 as  $(n_p - 1) = (10 - 1)$ ; for the  $pi$  effect, the degrees of freedom is  $(n_p - 1)(n_i - 1) = (10 - 1)(5 - 1) = 36$ , etc. The means are calculated by dividing the sum of squares for a given effect by the associated degrees of freedom. The obtained mean square must be converted into an expected mean square to reflect the universe of admissible observations. In other words, the expected mean square is the average mean square that would be obtained over infinite samples taken from the same population and universe (Shavelson & Webb, 1991).

Brennan (2010) details a procedure for estimating the variance components directly from mean squares using the following technique (p. 80):

1. A given effect ( $\alpha$ ) consists of  $t$  indices.
2. Let  $A$  = any component that consists of the indices contained in  $t$  plus one additional component
3. Apply the above to the formula:

$$\hat{\sigma}^2(\alpha) = \frac{1}{\pi(\alpha)} \left[ \begin{array}{l} \text{some combination} \\ \text{of mean squares} \end{array} \right]$$

where  $\pi(\alpha)$  = the product of the sample size for all G study indices not contained in ( $\alpha$ )

4. The combination of mean squares is defined as  $MS(\alpha)$



- a. Subtract the MS for any components that consist of  $t$  and one indices from  $A$
- b. Add MS for any components that consists of  $t$  and two indices from  $A$
- c. Continue adding components following the procedures for step a and b, stopping on the subtraction procedure if the total number of components is odd, and step b if even.

Applying these rules to the sample p x i x o design produces equation 4 below. The simplest method to calculate variance components for a mixed model is to specify all G study facets as random and then specify fixed facets in the D study design. As noted by Brennan (2010), this method produces unbiased estimates of true score and error variance.

However, Shavelson and Webb (1991) provide two alternate methods that yield more precise estimates. The first approach is to average the random components over conditions of the fixed facet. Alternatively, if the averaging approach does not make conceptual sense, each condition of the fixed facet may be run in a separate G study.

$$\begin{aligned}
 \hat{\sigma}^2(p) &= \frac{MS(p) - MS(pi) - MS(po) + MS(pio)}{n_i n_o} \\
 \hat{\sigma}^2(i) &= \frac{MS(i) - MS(pi) - MS(io) + MS(pio)}{n_p n_o} \\
 \hat{\sigma}^2(o) &= \frac{MS(o) - MS(po) - MS(io) + MS(pio)}{n_p n_i} \\
 \hat{\sigma}^2(pi) &= \frac{MS(pi) - MS(pio)}{n_o} \\
 \hat{\sigma}^2(po) &= \frac{MS(po) - MS(pio)}{n_i} \\
 \hat{\sigma}^2(io) &= \frac{MS(io) - MS(pio)}{n_p} \\
 \hat{\sigma}^2(pio) &= MS(pio)
 \end{aligned} \tag{4}$$

Estimated variance components are interpreted by their relative magnitudes. Thus the percentage of the total variance is reported for each component. The G study variance components reflect the error in generalizing from observed score variance to universe score variance using scores obtained by single person-item-occasion combinations. D study variance

components reflect the average score across persons and/or facets (Brennan, 2010; Webb et al., 2006).

**D study statistics.** A similar linear model can be used to represent the mean observed score over a given sample size of items and observations:

$$X_{pIO} = \mu + v_p + v_I + v_O + v_{pI} + v_{pO} + v_{IO} + v_{pIO} \quad (5)$$

The D study variance components are calculated using the components from the G study with the selected sample size for each facet, using the following rules described by Brennan (2010, p. 10):

$$\hat{\sigma}^2(\bar{\alpha}) = \begin{cases} \hat{\sigma}^2(\alpha)/n'_i & \text{if } \alpha \text{ contains } i \text{ but not } o \\ \hat{\sigma}^2(\alpha)/n'_o & \text{if } \alpha \text{ contains } o \text{ but not } i \text{ or} \\ \hat{\sigma}^2(\alpha)/(n'_i n'_o) & \text{if } \alpha \text{ contains both } i \text{ and } o \end{cases} \quad (6)$$

Using the following G study estimate variance components from a similar design in Brennan (2010, p. 8):

$$\begin{aligned} \hat{\sigma}^2(p) &= .25 & \hat{\sigma}^2(i) &= .06 & \hat{\sigma}^2(o) &= .02 & \hat{\sigma}^2(pi) &= .15 \\ \hat{\sigma}^2(po) &= .04 & \hat{\sigma}^2(io) &= .00 & \hat{\sigma}^2(pio) &= .12 \end{aligned}$$

A D study with the same p X I X O fully crossed random effects model with 8 participants, 6 items, and 4 assessment occasions, yields the following D study variance estimates:

$$\begin{aligned} \hat{\sigma}^2(p) &= .25 & \hat{\sigma}^2(I) &= .01 & \hat{\sigma}^2(O) &= .005 & \hat{\sigma}^2(pI) &= .025 \\ \hat{\sigma}^2(pO) &= .01 & \hat{\sigma}^2(IO) &= .00 & \hat{\sigma}^2(pIO) &= .005 \end{aligned}$$

Note that Brennan (2010) uses capital letters to distinguish D study variance components, a convention retained in the current study.

After the D study variance components have been calculated using desired sample sizes, the universe score variance and error variance are calculated. *Universe score variance* [ $\hat{\sigma}^2(\tau)$ ] is the variance of universe scores across all persons within the population. Analogous to true

score variance in CTT, universe score variance reflects true differences between individuals on the measured trait or behavior (Brennan, 2010).

Measurement error reflects inaccuracies in generalizing from an observed score to the universe score. Different variance components contribute to measurement error for absolute versus relative decisions (See Figure 1 and 2 for examples). *Relative error variance*,  $\hat{\sigma}^2(\delta)$ , is defined as the difference between the observed deviation score and the universe deviation score. Relative error variance includes components that influence relative ranking, namely those components that interact with persons (Brennan, 2010; Shavelson & Webb, 1991):

$$\hat{\sigma}^2(\delta) = \sigma^2_{pI} + \sigma^2_{pO} + \sigma^2_{pIO} \quad (7)$$

In our example with a sample of six items and four occasions,  $\hat{\sigma}^2(\delta) = .025 + .01 + .005 = .04$ .

*Absolute error variance* is the difference between the individual's observed score and universe scores (Brennan, 2010). Any component not contributing to universe score variance contributes to absolute error variance (Shavelson & Webb, 1991):

$$\hat{\sigma}^2(\Delta) = \sigma^2_I + \sigma^2_O + \sigma^2_{pI} + \sigma^2_{pO} + \sigma^2_{IO} + \sigma^2_{pIO} \quad (8)$$

In our example,  $\hat{\sigma}^2(\Delta) = .01 + .005 + .025 + .01 + .00 + .005 = .055$ .

After calculating the relative and absolute error variances, generalizability and dependability coefficients can be calculated. Similar to reliability in CTT, the generalizability coefficient is the ratio of universe score variance to the expected observed score variance (Webb et al., 2006):

$$\hat{\sigma}^2(\delta) = \sigma^2_{pI} + \sigma^2_{pO} + \sigma^2_{pIO} \quad (9)$$

In our example with a sample of six items and four occasions,  $\hat{\sigma}^2(\delta) = .025 + .01 + .005 = .04$ .

*Absolute error variance* is the difference between the individual's observed score and universe scores (Brennan, 2010). Any component not contributing to universe score variance contributes to absolute error variance (Shavelson & Webb, 1991):

$$\hat{\sigma}^2(\Delta) = \sigma_I^2 + \sigma_O^2 + \sigma_{pI}^2 + \sigma_{pO}^2 + \sigma_{IO}^2 + \sigma_{pIO}^2 \quad (10)$$

In our example,  $\hat{\sigma}^2(\Delta) = .01 + .005 + .025 + .01 + .00 + .005 = .055$ .

After calculating the relative and absolute error variances, generalizability and dependability coefficients can be calculated. Similar to reliability in CTT, the generalizability coefficient is the ratio of universe score variance to the expected observed score variance (Webb et al., 2006):

$$E\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)} \quad (11)$$

In our example with six items and four occasions,  $E\rho^2 = .86$ . Dependability is calculated for absolute decisions using the equation:

$$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)} \quad (12)$$

In our example,  $\Phi = .82$ .

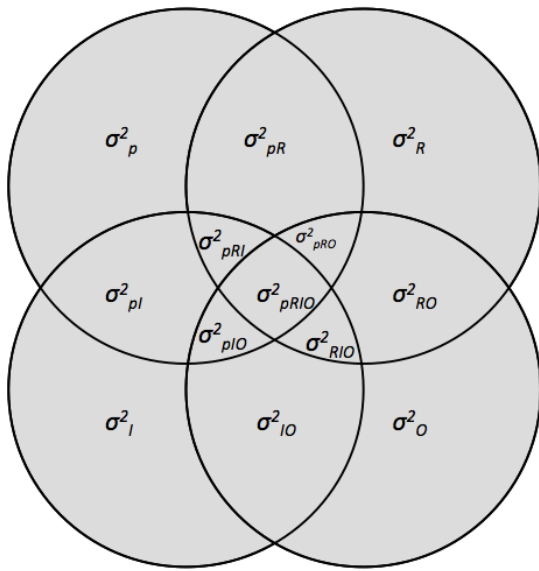
The sample size of the conditions can be adjusted in a follow-up D study. By adding two more occasions, the error variance is reduced to  $\hat{\sigma}^2(\delta) = .035$  and  $\hat{\sigma}^2(\Delta) = .048$ . As can be deduced from these calculations, new generalizability and dependability coefficients and SEMs are generated each time conditions are manipulated. In G theory, SEM is the square root of the sum of the estimated variance components that contribute to relative or absolute error variance (Brennan, 2010). As in CTT, the SEM can be used to create a confidence interval around an observed score. The universe score should fall within this range (Cronbach et al., 1972).

When including a fixed facet, the interaction between the fixed facet and the object of measurement enters into the universe score variance (Brennan, 2010). Thus, the variance due to true score increases and error variance decreases. This improves reliability (i.e., generalizability and dependability improves). This is made clear by comparing the ratio of true score to error variance depicted in the random model in Figure 1 to the Figure 2 model with items fixed. In our example, with items as a fixed facet in a D study with six items and six occasions:  $\hat{\sigma}^2(\delta) = .01$  and  $\hat{\sigma}^2(\Lambda) = .013$ , resulting in further reduction in error and improvements in generalizability and dependability.

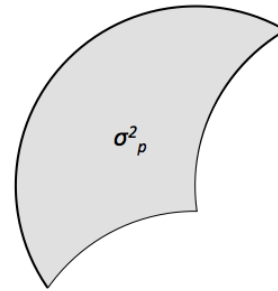
### **Relevance of G Theory to Behavioral Assessment**

G theory meets more of the behavioral assessment assumptions than CTT. G theory is based on a domain sampling approach in which observed measurements conditions (e.g., behaviors, items, raters, settings) are considered as a representative sample of similar conditions (Kane, 1982). As described by Gresham and Carey (1988), the basic assumption of the domain sampling approach as applied to G theory is that the means, variances and covariances of the obtained sample, if adequate, will reflect the means, variances and covariances of the larger universes. Applying this approach, the observations used to derive variance component estimates in G studies serve as starting point for estimating the reliability and error variances for a variety of measurement situations within D studies (Suen & Pui-Wa, 2007). This facilitates a flexible approach to assessment in which the measurement conditions do not need to be strictly set to determine reliability. In contrast, the parallel tests assumption of CTT requires identical conditions to examine reliability (Brennan, 2011).

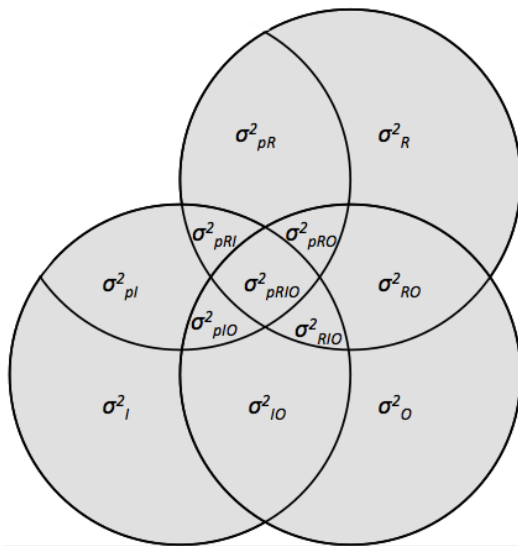
The primary advantage of such an approach is that the conditions of the measurement influence the reliability and error variance. Using D studies, the user can investigate how



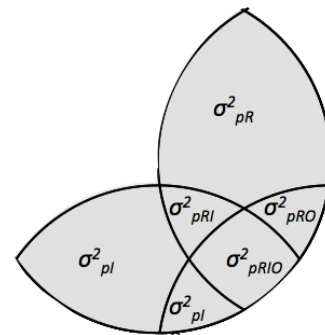
Total Variance for the Model



Universe Score Variance  
 $\sigma^2(\tau)$

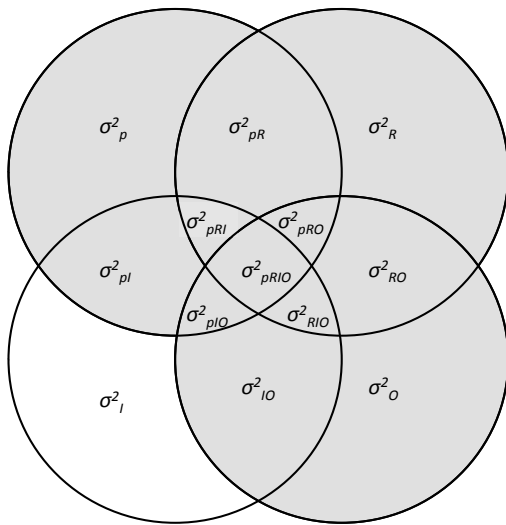


Absolute Error Variance  
 $\sigma^2(\delta)$

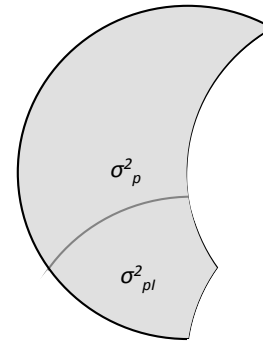


Relative Error Variance  
 $\sigma^2(\Delta)$

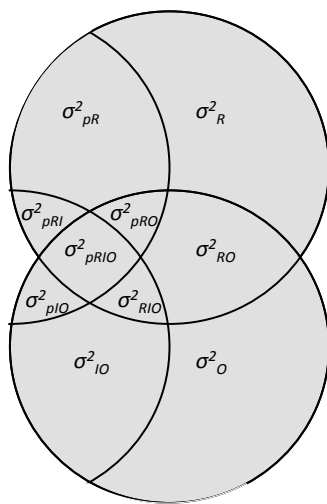
Figure 1. Venn Diagrams Displaying Sources of Variance For p X R X I X O Fully Crossed Random Effects Model Study Design.



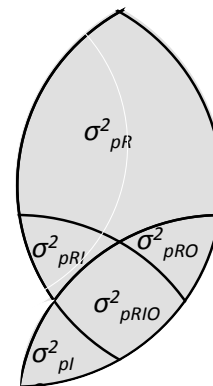
Total Variance for Model



Universe Score Variance  
 $\sigma^2(\tau)$



Absolute Error Variance  
 $\sigma^2(\delta)$



Relative Error Variance  
 $\sigma^2(\Delta)$

Figure 2. Venn Diagrams Displaying Sources of Variance For p X R X I X O Fully Crossed Mixed Model Study Design, with I Fixed and R and O Random.

changing these conditions will affect reliability (accuracy of decisions) without collecting additional data. This is quite similar to the assertion made by the behavioral assessment assumption that behavior is influenced by environment variables, e.g., setting, occasions, persons present, etc. (Chafouleas et al., 2007; Gresham & Carey, 1998). G theory is thus well suited to evaluate environmental factors that can influence behavior. For example, settings or raters can capture the situational specificity of behavior, while occasions may capture changes in behavior over time.

Moreover, Gresham and Lambros (1998) noted that G theory facilitates a best practice approach to behavioral assessment. Behavior should be assessed across multiple settings, occasions, therapists, and methods (also see Cone, 1978). Collecting behavioral assessment in this manner increases the generalizability of the assessment data to other setting, or situations. In fact, consideration of the generality of effects across various settings and conditions is one of the defining characteristics of applied behavior analysis (Baer, Wolf, & Risley, 1968; 1987).

Another assumption of G theory is that the behavior being measured occurs within steady state responding. Differences between individuals are not due to systematic changes in learning or maturation (Shavelson & Webb, 1991). This assumption is also reflected in behavioral assessment, as steady-state responding in the absence of intervention is an underlying principle of single-case design logic (Kazdin, 1982). However, this assumption also reflects a limitation of G theory. It cannot account for changes due to intervention. If changes due to intervention did occur within a G study, the error due to facet of occasions could be quite large, detracted from detection of individual differences.

Volpe, Briesch & Gadow (2011) noted that there are two separate but equally important questions to be addressed when a measure is developed for progress monitoring purposes: 1) is



the measure reliable?, which can be addressed by G theory; and 2) is the measure sensitive to changes in behavior?, which is typically addressed in a separate study. This limitation notwithstanding, G theory is able to account for both relative and absolute decisions, with the former being important for screening decisions and the latter being relevant in a progress monitoring context when within-individual decisions are made using a criterion or self-referenced approach (Christ & Hintze, 2007).

A further advantage of G theory is that it can be used to design more efficient and cost-effective measurement procedures (Parkes, 2000; Volpe et al., 2011). Namely, D studies can be used to derive various combinations of measurement conditions that will each achieve acceptable levels of generalizability or dependability. As detailed by Volpe and colleagues, the advantages of this approach are as follows: 1) by quantifying the cost associated with each condition (e.g., number of raters or administrations, level of training required, and financial considerations) the various combinations could be ranked by efficiency and cost; 2) a menu of assessment options can be developed, which may increase assessment acceptability; 3) the purposes and constraints of the assessment system can be used to select an appropriate option. For example, if a decision needs to be made quickly, an option with more items and raters can be selected.

## **REVIEW OF METHODS USED TO ASSESS SOCIAL BEHAVIOR**

This section will review several methods currently used to assess social behavior.

Reviews will include an overview of the method, advantages and disadvantages, supporting evidence for psychometric properties as evaluated by CTT and G theory, and an evaluation of the feasibility of the method for progress monitoring use within an RTI context.

### **Systematic Direct Observation**

Direct observation is widely used to assess behavior (Hintze, Volpe, & Shapiro, 2008; Wilson & Reschly, 1996) and has been considered as the “gold standard” for formative assessment (Volpe, DiPerna, Hintze, & Shapiro, 2005, p. 471). Defining characteristics of SDO include specification of operationally defined behaviors, standardization of observation procedures, consideration of contextual factors (e.g. time and setting where the observation is conducted) and standardization of scoring procedures to further reduce unwanted variability across observers (Salvia, Ysseldyke, & Bolt, 2004).

There are several advantages in using SDO to assess social behavior. First, SDO is a direct and objective measure of behavior. Next, target behavior can be defined to reflect broad or specific target behaviors. Importantly, SDO is sensitive to detecting changes in behavior and used to evaluate intervention effects (Hintze, et al., 2008; Volpe & Gadow, 2010). Finally SDO can assess different dimensions of behavior such as frequency (event recording or time sampling methods), temporality (duration, latency or inter-response time; Alberto & Troutman, 2003).

However, several disadvantages limit the feasibility of SDO as a progress-monitoring tool. First, SDOs are typically conducted by an external observer, such as school psychologists with advanced training in observational methods (Riley-Tillman, Chafouleas, Briesch & Eckert, 2008). Next, 5 to 40 fifteen-minute observations may be required for accurate assessment

(Briesch et al., 2010; Hintze & Mathews, 2004). Thus, SDO requires a significant investment of time and resources. Furthermore, SDO is inappropriate to measure behaviors that occur infrequently (Saudargras & Zanolli, 1990). SDOs may also induce observer reactivity, which may lower the acceptability of its use (Merrell, 2003). Finally, given the situational specificity of behavior, observations may not generalize across settings (Shapiro, 1988).

**Psychometric properties.** As SDO follows the assumptions of behavioral assessment, attempts to establish the psychometric properties of SDO using CTT have been inconclusive. The resulting ideological differences have led to alternate conceptualizations of reliability and validity (Hintze, 2005). Baer (1977) defined reliability as the degree to which independent observers agree on the occurrence or nonoccurrence of a behavior, referred to as *interobserver agreement* (Cooper et al, 2007; Kazdin, 1982). As noted by Gresham (2003), IOA has long been used as the standard for reliability by the field of applied behavior analysis. Using this definition, the reliability of direct observation is thus well established.

In SDO, content validity is defined as the degree to which the observation is representative of behavior under all relevant conditions (Gresham, 2003). Criterion-related validity and content-related validity are similar to the traditional definitions (Hintze, 2005). Some behaviorists have replaced traditional validity with accuracy, defined as the degree to which the observed behavior accurately reflects the true value, or actual occurrence of behavior (Johnson & Pennypecker, 1993). As described by Cone (1988, as cited in Hintze & Mathews, 2004), an accurate measurement system must reflect the true occurrence of the behavior over multiple occasions and settings. Hintze and Mathews noted this might be accomplished best by using G theory.

**Generalizability of direct observations.** Two studies have investigated the generalizability and dependability of direct observation with mixed results. Hintze and Matthews (2004) compared the generalizability of SDOs across setting and times with interobserver agreement. In this study, 14 fifth-grade students were observed twice a day for 10 consecutive school days. Five graduate students conducted the observations. On-task and off-task behavior were coded using a 15-second momentary monetary time sampling procedure. These procedures yielded a 14 students (person) x 2 settings (math and ELA) x 10 occasions (10 consecutive school days) fully crossed G study design.

As expected, nearly two-thirds of the variance was accounted for by person, which reflected individual differences in on-task/off-task behavior. Other significant variance was attributed to the person X setting interaction (13%) and the residual (24%). With the initial model, generalizability and dependability coefficients were low ( $E\rho^2 = .63$  and  $\Phi = .62$ ). A series of D studies was conducted to determine how many observations were required to achieve adequate generalizability and dependability. Acceptable levels were only obtained if students were observed four times a day for 40 days. However, for some students, this could be obtained after 7- 20 observations (Volpe, McConaughy, & Hintze, 2009). In contrast, IOA was high, averaging 90%. This suggested that IOA should not be used as a proxy for reliability, an assertion made by others previously (Kazdin, 1982).

In 2010, Briesch, Chafouleas, and Riley-Tillman conducted a study that compared SDO and DBR. Participants were 12 kindergarten students in an inclusive classroom, their two teachers and two graduate research assistants. Students were observed daily during a 45-minute instructional block for 10 consecutive school days. After each 15-minute period within this block, teachers completed DBRs for academic engagement. The research assistants reviewed

videos for each period and coded for academic engagement using a 15-second momentary time sampling procedure. Four students were observed at a time. SDO data were converted into a similar metric as the DBR scale to facilitate comparisons.

These procedures produced 12 person X 2 rater X (3 occasions nested into 10 days) random effects design. Separate G and D studies were conducted for each method (see DBR section for DBR results). For SDO, person accounted for 48% of the variance, person X occasion: day accounted for 30% of the variance, and residual accounted for 14%.

Generalizability and dependability coefficients for the enacted conditions were high ( $E\rho^2 = .98$  and  $\Phi = .97$ ), but this model specified 30 observations, which is unrealistic in school settings. Therefore, follow-up D studies were conducted to examine the effect of using one observer and one observation per day. Generalizability and dependability were adequate for SDO after 5 days,  $E\rho^2 = .83$  and  $\Phi = .82$ , but never reached acceptable levels for DBR. For SDO, adequate reliability was achieved more efficiently than reported in the Hintze and Mathews (2004) study, even with several students observed simultaneously. The authors suggested that discrepant findings might be due to differences in observer training, operational definitions, the addition of the rater facet (i.e., better specification), or the use of analogue observations.

**Summary.** Although considered as a “gold standard” formative assessment tool, SDO has limited utility as a progress-monitoring tool within an RTI context (Riley-Tillman, et al., p. 119; Volpe & Gadow, 2010). Chief among these limitations are the resources required to obtain reliable behavior observations. If at minimum five 15-minute observations are required, this time commitment would be considerable when conducted for all Tier 2 or Tier 3 students within a school. In summary, SDO is perhaps best suited for making high-stakes decisions where time is not an issue (Briesch & Volpe, 2007).

## Behavior Rating Scales

Behavior rating scales (BRS) are questionnaires or checklists in which informants familiar with the student (e.g., parents, teachers, or the student) rate the frequency of relevant behaviors over time (Busse, 2005; Elliott, Busse & Gresham, 1993). This is an indirect form of assessment, as the behavior is not measured at the time and place at which it occurs (Briesch & Volpe, 2007). Items on BRS assess specific behaviors that are summated to represent clusters of related behaviors or symptoms. Depending on the scope of dimensions measured, BRS are classified as global or narrow band (Merrell, 2003).

Several assumptions guide the use of rating scales. First, as ratings involve evaluative judgments, raters may have different standards for behavior (Busse, 2005). This may also reflect different expectations for behavior across settings. For example, horseplay may be acceptable at recess, but not during class. A consistent finding across rating scales is that correlations between informants in same settings yield higher levels of agreement than informants in different settings (Achenbach, McConaughy, & Howell, 1987; De Los Reyes, & Kazdin, 2005; Gresham, Elliott, Cook, Vance, & Kettler, 2010). Finally, BRS reflect summative evaluations of behavior over time. Thus, ratings indicate the perceived frequency rather than the absolute frequency of behavior (Elliott, Busse & Gresham, 1993).

Behavior rating scales offer some important advantages as an assessment tool. First, BRS allow for ease in quantifying behavior. They may be completed by multiple raters, which facilitates comparisons of behavior across settings and informants as part of a multisource, multisetting, multimethod assessment (Busse, 2005; Merrell, 2003). Next, BRS can capture intense behaviors that occur at a low frequency (Merrell). Behavior ratings scales are cost-effective, and require little training or resources as they are completed by informants who

interact with the child in natural contexts (Volpe, Gadow, Blom-Hoffman, & Feinberg, 2009). Additionally, BRS may be either norm-referenced or criterion referenced (Merrell; Wilson & Reschly, 1996). Finally, many BRS with excellent technical features are available for a wide range of behavioral domains.

However, there are several disadvantages. First, ratings may not be sensitive to short-term treatment effects (Christ et al., 2009). It can also be difficult to reconcile conflicting ratings from different informants (Kraemer, et al., 2003). While discrepant ratings may indicate the situationally specific behavior, BRS do not indicate the specific context in which behavior occurs (De Los Reyes & Kazdin, 2005). For example, disruptive behavior may occur more during math than reading, which may have important treatment implications (Christ et al.). Lastly, as ratings reflect evaluative judgments, rater biases may emerge. Common sources of rating error include halo effects, leniency errors, and central tendency errors (Merrell, 2003).

Additional limitations arise in the use of BRS as a progress monitoring tool. As noted by Volpe et al. (2011), most BRS are not designed for repeat administration within a short time frame. From a practical perspective, the length of BRS limits the acceptability for frequent administrations (Volpe et al., 2009; Volpe, Heick, & Guerasko-Moore, 2005). While there are a few abbreviated BRS, such as the ADHD Symptom Checklist (Gadow & Sprafkin, 2008) that are designed for frequent use, the length of these scales still would prohibit the use in an RTI context in which teachers would rate about 15% of students (Volpe & Gadow, 2010).

**Psychometric properties.** Although the specific psychometrics depends on the quality of the rating scale, there are many established behavior ratings scales with excellent reliability and validity. Broadband examples include the Achenbach Child Behavior Checklist scales (ASEBA, Achenbach & Rescorla, 2001) and the Behavior Assessment System for Children

(BASC, Reynolds & Kamphaus, 2004), while narrow-band examples include the NCIHQ Vanderbilt scales for ADHD (Wolraich, Feurer, Hannah, Pinnock, & Baumgaertel, 1998) and Social Skill Improvement System Rating Scales (Gresham & Elliott, 2008).

**Generalizability of BRS.** To date, the psychometric features of BRS have been almost universally established via CTT. Despite the relevance of G theory to behavioral assessment- particularly in examining error associated with raters and settings- only one BRS G theory study could be located. In this study, Bergeron, Floyd, McCormack, and Farmer (2008) explored the generalizability and dependability of externalizing behavior scales on the BASC and ASEBA Teacher Report Form. Participants were 61 elementary school students within six classrooms. Six teacher dyads rated each student in their respective classrooms on the BASC and ASEBA on two occasions over several weeks. This represented a partially nested design with 6 teachers and 61 students nested in 6 classrooms, fully crossed with 2 occasions and 2 instruments. Thus, main sources of variance included classroom, student within classroom, rater within classroom, occasions and instrument, and interactions terms included student by rater within classroom, classroom by occasion, classroom by instrument, occasion by instrument, and residual error.

As expected, students within classroom accounted the largest portion of variance in the externalizing composite scores (67.7%). The residual accounted for 11.5% of the variance, student by rater within classroom for 9%, rater within classroom for 4.7%, instrument for 2.4% and classroom by occasion for 2.1%. Dependability was .68 for the externalizing composite, which is below acceptable levels for progress monitoring decisions.

Additionally, traditional reliability analyses were calculated using Pearson correlations:  $r = .89$  across instruments, test-retest was .93 and .89 for the two measures, and correlations between the raters were .79 and .73. This finding highlights our earlier assertion that traditional



reliability estimates may not accurately account for all sources of error in authentic testing situations.

**Summary.** Behavior ratings scales have long been used to classify, diagnosis and evaluate long-term treatment outcomes for social behavior. They are easy to use, enjoy a long history of psychometric support for well-constructed scales, and require relatively few resources. However, BRS are inappropriate for progress monitoring purposes for several reasons. First, BRS are not designed to detect short-term intervention effects. Next, the time required to complete lengthy scales makes frequent assessment impractical, particularly when assessing several students. Finally, an initial G theory study of behavior ratings scales calls into question the reliability of behavior rating scales when several sources of error are considered simultaneously.

### **Direct Behavior Ratings**

Direct Behavior Ratings have been proposed as a viable progress monitoring tool (Chafouleas, et al. 2007; Chafouleas, Riley-Tillman, Sassu, LaFrance, & Patwa, 2007; Volpe & Briesch, 2012). DBRs are a hybrid assessment tool that blends features of systematic direct observation (SDO) and behavioral rating scales (Riley-Tillman, Christ, Chafouleas, Boice-Mallach & Briesch, 2010). After a specified interval of observation, teachers complete a rating of the estimated the frequency of target behaviors (Chafouleas, McDougal, Riley-Tillman, Panahon, & Hilt, 2005). Defining characteristics of DBRs include: (a) observation of specified behaviors by important consumers (e.g., teachers), (b) rating of behavior by these consumers during or directly following a set observation interval, (c) communication of results across consumers (parents, teachers, students), and (d) frequent collection of data for formative purposes (Riley-Tillman, Chafouleas, & Briesch, 2007; Riley-Tillman, Chafouleas, Christ,

Briesch, & LeBel, 2009). Although the use of DBRs as a progress monitoring tool has emerged within the past decade, DBR-like methods have a long history as intervention tools such as daily behavior report cards and good behavior notes (Chafouleas, Riley-Tillman & McDougal, 2002; Lahey, Gendrich, Gendrich, Schnelle, Gant, & McNees, 1977; McCain & Kelley, 1993).

As a hybrid between SDO and BRS, DBRs offer advantages of both methods. Like SDO, DBRs are a direct measure of behavior, and are designed for repeated administrations, which is an ideal feature for a progress monitoring tool. Like BRS, DBRs are typically completed by classroom teachers, which minimize reactivity concerns and the need for extensive training (Christ et al., 2009; Riley-Tillman et al., 2010). When longer rating intervals are used, DBRs may capture low frequency behaviors otherwise missed by SDO (Volpe & Chafouleas, 2011).

Furthermore, DBRs are a flexible tool that can be customized in terms of the target behavior selected, frequency of rating, rater and target of rating, and other features (Chafouleas, et al., 2002; Christ et al., 2009). A recent line of G theory studies suggested that DBRS can be developed with different number of gradients, or continuous versus discrete scaling (Briesch, et al., 2012; Chafouleas, et al., 2009; Christ, et al., 2010). Next, teachers generally find DBRs to be acceptable and many have experience with DBRs or similar tools (Chafouleas, et al. 2006). Finally, initial studies suggested that DBRs are sensitive to measuring effects of psychopharmacological and behavioral interventions (Chafouleas, et al., 2012; Pelham et al., 2002; Pelham, et al., 2005).

However, there are several disadvantages of DBR. First, as an emerging method, more evidence is needed to determine the technical adequacy of different target behaviors and rating formats. Here, the flexibility of DBRs may also prove somewhat of hindrance, as different target behaviors, populations and construction methods may influence the psychometric properties.

However, this limitation is not unique to this method. Next, like BRS, the DBR method is subject to rater biases. As will be described, G theory studies have generally shown more variability associated with the rater facets when ratings are completed by teachers, rather than trained observers. Finally, G theory findings suggested that at least two or more weeks of data should be collected to make accurate decisions. As noted by Volpe and Briesch (2012), this may be too long to wait to make intervention decisions in some cases.

**Psychometrics properties.** The specific psychometrics of DBR depend on the format and target items used. In one study, Fabiano and colleagues investigated the psychometric of a multi-item DBR used as a daily behavior report card (Fabiano, Vujnovic, Naylor, Parsieau & Robins, 2009). Items on the DBR were relevant to ADHD such as task completion and following directions. Test retest reliability (calculated by comparing even and odd days) was excellent ( $r = .94$ ). Interrater reliability was moderate ( $r = .46$ ) as calculated with correlations between the teachers' DBR scores and an external observer. Another study by Riley-Tillman et al. (2011) compared the test-retest reliability of 10 and 20 minute rating intervals of engaged behavior after one week. Test retest correlations were low to moderate (range = .31-56) for 10-minute rating interval conditions, and low to high (range = .31-1.0) for 20-minute rating interval conditions. Finally, DBRs show moderate to high correlations with direct behavior ratings (Chafouleas et al., 2005; Riley-Tillman, Chafouleas, Sassu, K.A., Chanese, & Glazer, 2008), and moderate correlations with the Social Skills Rating Scale when used as screener for social behavior (Chafouleas, Kilgus & Hernandez, 2009).

**Generalizability of DBRs.** Unlike other methods, the reliability of DBR has been predominately established using G theory. In the first study in this series, Chafouleas and colleagues investigated the generalizability and dependability of the DBR items: “works to

resolve conflicts with peers” (WRC) and “interacts cooperatively with peers” (IC, Chafouleas et al., 2007, p. 68). Participants were 15 students in a university-affiliated preschool classroom and four of their classroom teachers. The teachers rated each student twice a day for 13 consecutive school days after 30-minute intervals during small group instruction or free play. Ratings were completed by placing an X on a continuous line with 3 anchors at 0, 50 and 100% to indicate the proportion of the interval that students engaged in the target behavior. These procedures yielded a 15 person by 2 setting by 4 rater by 13 occasion fully crossed random effects design.

Separate G and D studies were conducted for each of the DBR items. For WRC, the rater facet accounted for 41% of the total variance, followed by the residual error (23%), person (18%), person by rater (8%), and rater by day (5%). Interestingly, WRC was more influenced by rater effects than for IC. For IC, rater effects were still significant (28% of the total variance). The highest proportion of variance was accounted for by person at 38% of the total variance, and 27% of the variance was unexplained. Reduced G studies were conducted for each rater and behavior, while observations were collapsed which yielded a 15 person by 26 observations fully crossed design. The true score variance improved substantially, with 30-60% of the variance attributed to person, 4-7% attributed to occasion and 34-64% attributed to the residual.

A series of D studies evaluated the number of occasions needed to obtain adequate reliability for progress monitoring purposes. Generalizability and dependability reached acceptable levels for low stakes decision ( $\geq .70$ ) after 4-7 ratings, and for high stakes decisions ( $\geq .90$ ) after 10 ratings (Chafouleas et al., 2007). Furthermore, Chafouleas et al. (2007) noted that given the brief observation interval, the teachers may not have interacted with the students equally while carrying out their normal teaching duties. Given these findings, the authors suggested that DBR outcomes should be evaluated with one consistent rater.

Next, Christ et al. (2010) examined the generalizability and dependability of two different single-items DBR under tightly controlled conditions to examine the effect of various assessment procedures and interpretive assumptions (e.g., effect of fixed vs. random facets) with an emphasis in evaluating the effects of raters. In this study, 125 undergraduate research participants observed 1-minute video clips of 27 children working on an unsolvable task with Legos. The two DBR items rated were Visually Distracted and Actively Manipulating. Each participant rated 18 video clips three times using different scaling conditions. This resulted in a 27 person by 125 rater by 3 occasion design. Multiple series of G and D studies were conducted separately for each DBR item and each condition.

The study had several interesting findings. First, the variance was portioned almost identically across the three scaling conditions, which indicated flexibility of scaling options. Next, variance due to persons was much lower than in previous studies. Similar to Chafouleas et al., 2007, 21% of the variance for both DBR items was accounted for by rater facet, and 11% for person by rater, which suggested that DBRs may need to be interpreted by rater. Finally, the universe interpretation greatly affected the amount of occasions and raters necessary to make reliable decisions. In the restricted universe, adequate generalizability was obtained with one rater after 5 occasions for low stakes decisions (.70) and 15-20 occasions for high stakes decisions (.90). Dependability required 8-10 occasions to reach .70, and more than 20 occasions for .90. Adding additional raters did not significantly impact these estimates. In contrast, 15-20 observations with three raters were needed to obtain .70 generalizability within the infinite universe. Adequate dependability was never achieved, even when with 5 raters and 20 occasions. This highlights the need to consider how the outcomes will be used.

In a follow-up to the Chafouleas et al. 2007 study, Chafouleas et al. (2010) investigated the generalizability and dependability of DBRs for academic engagement and disruptive behavior. Participants were six students in an inclusive middle school classroom, their classroom and a consultant teacher, and two research assistant observers. The study was conducted over six consecutive school days during English and language arts instruction. Each ELA period was divided into 10-minute observation intervals, during which the teachers conducted normal teaching duties while the research assistants observed the class. At the end of each interval, each rater completed the DBRs for each student. These procedures led to a 6 person by 4 rater by 3 occasion: 6 day partially nested design.

In contrast to previous studies, person accounted for only 23-25% of the variance; however person by day accounted for 14-18%, which indicated some variability in behavior across days. Surprisingly, rater effects accounted for only 2-5% of the variance, and interactions between raters and other facets were negligible. Notably, 40-43% of the variance was unexplained in the initial model, but was reduced to 35-40% when calculated by rater. The head teacher was found to have similar ratings to the two research assistants. Visual inspection of the ratings indicated disagreement on the absolute level of students' behavior, but agreement on the relative ranking of behavior between students. For the two research assistants and the head teacher, both generalizability and dependability reached acceptable levels (mean  $E\rho^2$  of .87 and  $\Phi$  of .76) with three ratings per day for six days. Alternatively, ratings could be conducted once a day for 10-15 days to reach for .80, or for 15-20 days for .90 for high-stakes decisions. For the consultant teacher, more than 20 days of data were required to obtain  $E\rho^2$  of .78 and  $\Phi$  of .73. The authors suggest that the discrepant ratings between the two teachers may reflect different

amounts of time spend with students; the consultant teacher may have interacted with only part of the class during the observation intervals (Chafouleas et al., 2010).

The G study conducted by Briesch et al. (2010) comparing the generalizability and dependability of SDO and DBR was previously discussed in the SDO section. As a brief review, two kindergarten teachers rated 12 students on a DBR for academic behavior three times a day for 10 days. Person accounted for 47% of the variance, person by rater accounted for 20%, person by rater for 7.5%, and the residual three-way interaction accounted for 13%. For the initial model with 30 ratings and two teachers, generalizability and dependability were  $E\rho^2 = .82$  and  $\Phi = .77$ . However, using one teacher 20 days of DBRs would need to be collected to reach  $E\rho^2 = .70$  and  $\Phi = .63$ , and dependability was never reached. In contrast to Chafouleas et al. 2010, increasing the number of observations per day had little effect. One possible explanation for these discrepant findings is that the teachers rated more students during the same length of time, and may not have observed or interacted each student as when fewer students were rated.

**Multiple item scales.** The research of Chafouleas, Riley-Tillman, Christ and have colleagues have predominately focused on single-item DBR scales. As noted by Christ et al. (2009), the primary aim of this research group is to establish a series of DBR single item scales that can be used as GOMs. Alternatively, DBRs may be constructed with multiple items to assess specific target behaviors as an ideographic approach to progress monitoring, as was previously described in the review of the Fabiano et al. (2009) study.

A recent study by Volpe and Briesch (2012) highlights an important advantage of multiple item scales- fewer occasions may be required to obtain adequate reliability. In this study, the authors compared the generalizability and dependability of DBR single item scales

(SIS) to multiple-item scales (MIS) for academic engagement and disruptive behavior. The MIS consisted of five items describing specific behaviors related to the construct (e.g., “out of seat” and “calls out”) whereas the SIS consisted of one item describing the global construct (e.g., disruptive behavior), with the corresponding MIS items provided as examples (Volpe & Briesch, p. 246). Participants were eight middle school students, videotaped during their math class for three days. Two doctoral students observed 10-minute video clips, rating one student at a time. These features were counterbalanced to control for order effects so that each rater could complete both types of ratings for all students across all observations.

These procedures resulted in an 8 person by 2 rater by 3 occasion fully crossed, random effects model design. G and D studies were conducted separately for each method and construct. Across both academic and disruptive behavior, a higher proportion of the variance was attributed to person for the MIS scale than on the SIS (67% MIS vs. 46% SIS for academic; 37% MIS vs. 29% SIS for disruptive behavior), which indicated that MIS was more sensitive to detecting individual differences in behavior. Person by occasion accounted for 18% of the variance for MIS, and 25% for SIS for academic engagement, and for 31% for MIS and 49% for SIS for disruptive behavior, indicated that students’ behavior was less consistent for disruptive behavior than for academic engagement. Notably, rater effects and residual variance were lower than previous studies across both behaviors and types of scales.

Across all D studies, fewer occasions were necessary to obtain acceptable levels for MIS than for SIS. For the initial model with two raters and three occasions, generalizability and dependability for academic engagement were  $E\rho^2 = .85$  and  $\Phi = .82$  for MIS and  $E\rho^2 = .73$  and  $\Phi = .70$  for SIS and  $E\rho^2 = .64$  and  $\Phi = .63$  for MIS compared to  $E\rho^2 = .50$  and  $\Phi = .49$  SIS for disruptive behavior. In order to reach the .80 standard with one



rater, only two occasions were needed for relative decisions, and four occasions for absolute decisions for academic motivation, compared to eight and seventeen occasions with DBR SIS. For disruptive behavior, 11 occasions were needed for relative decisions and 12 for MIS, whereas using SIS, more than 20 occasions were needed to reach .70, and .80 was never obtained even after 100 occasions. Although further research is needed to replicate these results, these findings suggested that the efficacy of DBRs can be improved by using a multiple item scale.

### **Brief Behavior Rating Scales**

Recently, brief behavior ratings scales (BBRS) have been proposed as an alternative progress monitoring tool (Gresham et al., 2010; Volpe, et al., 2009). BBRS contain a subset of items from established BRS (Volpe et al., 2011). Several abbreviated versions of established rating scales are currently available, such as the BASC Monitor for ADHD (Reynolds & Kamphaus, 2004), the IOWA Conners Teacher Rating Scale (Looney & Milich, 1982), the Revised Children's Manifest Anxiety Scale (Reynolds, 2008), and the Children's Depression Inventory-Short Form (Kovacs, 2011).

Until recently, use of BBRS for progress monitoring was mostly restricted to well-funded research (e.g., ADHD medication titration studies; Gadow, Nolan, Paolicelli & Sprafkin, 1991; Pelham et al., 2002; Power, DuPaul, Shapiro, & Kazak, 2003). However, most of these abbreviated rating scales still have 30 or more items, which limits progress monitoring within an RTI framework (Gureasko-Moore et al., 2005; Volpe et al. 2011). Most BBRS are designed to assess specific behavior or performance objectives. A notable exception is a change-sensitive abbreviated version of the Social Skills Rating Scale (Gresham & Elliott, 1990), discussed in detail shortly (Gresham et al., 2010).

There are several methods for constructing brief behavior rating scales. Nomothetic approaches, such as the factor-analysis and change-sensitivity methods, are developed with data derived from large samples, whereas in the ideographic approaches, custom scales are developed for each student. In the factor analysis method, a factor analysis is run on a large dataset of ratings and the items with the highest factor loadings are retained. As noted by (Volpe & Gadow, 2010), this approach should produce highly intercorrelated items that reflect the construct of interest, resulting in similar psychometric properties to the original scale.

The change-sensitivity method uses items that are demonstrated to be sensitive in detecting behavioral change in response to an intervention or treatment. In this approach, ratings are completed on the full-scale measure for a large dataset of students before and after an intervention. Several change-sensitivity metrics, such as an odds-ratio, standardized mean effect size, and reliable change index (Jacobson & Truax, 1991) are calculated and decision rules are applied to retain change-sensitive items (Gresham et al., 2010; Volpe & Gadow, 2010). This has clear advantages for progress monitoring, as these scales may be better at capturing treatment effects (Meier, McDougal, & Bardos, 2008). Both methods result in a standardized BBRS, which facilitates evaluation of the psychometric properties of the scales. However, the tradeoff in using standardized scales is that they may not contain the most relevant items for individual students (Volpe et al., 2009).

Hyman et al. (1998) advocated for a menu-based approach in which raters select relevant items from full-scale measures for inclusion on the abbreviated scale. These items are converted to a common Likert scaling and rated daily. Alternatively, customized scales can be developed by administering the full measure and then selecting the problematic items rated as most problematic for an individual (Gureasko-Moore et al., 2005; Volpe et al., 2009). Across both

methods, the resulting scales consist of items that are highly relevant for individual students. This may result in scales that are highly sensitive to detecting intervention effects. Furthermore, soliciting input from raters may increase acceptability and use of assessment procedures (Hyman et al). However, the psychometric proprieties of scales are difficult to evaluate because each scale consists of different items.

**Psychometric properties.** Support for the technical adequacy of BBRS is emerging. Recent research conducted by Volpe and colleagues (2009; 2010; 2011), and by Gresham et al. (2010) suggested that BBRS may retain similar psychometric features of the original full-length scales. First, Volpe et al. (2009) developed 4-item BBRS for two 9-item subscales (Inattentive and Hyperactive-Impulsive) of the ADHD-Symptom Checklist-4 (ADHD-SC4) with factor-derived and customized construction methods. This study used a dataset of ADHD-SC4 ratings completed by participants' teachers within a larger ADHD treatment study. Although the authors hypothesized that the abbreviated scales would be less reliable than full scales, the coefficient alpha was actually highest for the individualized BBRS ( $\alpha = .93, .87$  for full item scale,  $.72$ ) for the factor-derived scale on the Inattentive scales and comparable across the factor-derived, individualized and full scale versions ( $.93, .96$  and  $.96$  respectively) on the Hyperactive-Impulsive scales. Mean test-rest correlation coefficients were similar across the various versions of the scales. Finally, a multivariate analysis was conducted across baseline and treatment conditions to examine the change-sensitivity of the scales. Again, all three methods yielded similar results.

These results were replicated in a follow-up study (Volpe & Gadow, 2010). Factor-derived and individualized BBRS were constructed for three subscales (the 5-item IO and AG scales from the IOWA Conners and the 10-item Peer Conflict scale; Gadow, 1986), and

compared in term of reliability, treatment sensitivity, and concurrent validity with a methodology similar to the previous study. Across all conditions, and two out of the three scales, coefficient alphas were similar across the different versions of the scales ( $\alpha = .71$ -.91 factor derived scales, .80-.93 individualized scales, and .85-.93 full-version scales). The test-retest correlation coefficients were also comparable (.60-.81 factor derived, .65-.85 individualized, and .64-.81 full-version scales). The scales were compared to SDO and to the Child Symptom Inventory (Gadow & Sprafkin, 2002). While the overall findings for concurrent validity were mixed, the average correlation coefficients obtained for each method were similar. Finally treatment sensitivity was calculated for each method using a multivariate analysis approach. The individualized scale was the most sensitive, followed by the factor derived scale and then the full scale. Together with the Volpe et al. 2009 study, these findings indicate that that abbreviated ratings scales may retain or even improve technical features even though fewer items are administered.

The study by Gresham et al. (2010) also demonstrated that a BBRS could retain adequate psychometric features. The purpose of this study was to develop a change-sensitive BBRS of the Social Skills Rating Scale (SSRS, Gresham & Elliott, 1990) as GOM for social behavior while retaining acceptable psychometric properties. This study used a dataset of pre and post scores on the SSRS for 200 students who participated in a randomized clinical trial of the First Step to Success (Walker et al., 2009). Four change-sensitive metrics (odds ratio, standardized mean effect size, independent and dependent *t* tests, and ANOVA interaction effects) were calculated to develop a pool of items that were sensitive to change across pre and post-test scores. Items that demonstrated sensitivity across at least 3 of the four change-sensitivity metrics were retained.

Next, the technical adequacy of the resulting 29-item pool was assessed and used in an iterative process to develop the final BBRS. Items were ranked in order of change-sensitivity and the least sensitive item was dropped. Technical adequacy was recalculated and the item with the next lowest sensitivity was then considered for deletion. This process continued until the internal consistency and test-retest stability coefficients dropped below the .70 (a standard identified as acceptable for early scale development by Nunnally & Bernstein, 1994) and/or correlations with the TRF (Achenbach & Rescorla, 2001) and SSRS- Social Skills and Problem Behavior subscales were no longer considered strong (e.g.,  $r < .50$ ). This resulted in a 12-item BBRS with an internal consistency of  $\alpha = .70$ , three-month test-retest stability of .71 (for the control group), and criterion validity of  $r = .51$ ,  $-.59$ , and  $.54$  with the TRF, Social Skills and Problem Behavior scales respectively. The 12-item BBRS consisted of 8 items from Social Skill scale, 3 items from the Problem Behavior scale and 1 item from the academic competence scale, thus representing several constructs important to conceptualizations of social behavior (Gresham et al., 2010).

**Generalizability of Brief Behavior Rating Scales.** Only one G theory study of BBRS could be located. In this study, Volpe, Briesch, and Gadow (2011) investigated the generalizability and dependability of the IOWA Conners Teachers Rating Scale, with a focus on the number of occasions and items necessary to obtain acceptable reliability for decisions. This study used rating scale data obtained for 67 children diagnosed with ADHD (aged 6-13) that was collected as a part of a randomized clinical trial of a methylphenidate medication. The students were rated by their teachers four times in a two-week period on the IOWA Conners. The IOWA Conners scale contains two 5-item subscales: the inattentive-overactivity (IO) scale, and the

oppositional-defiant (OD) scale. These procedures yielded a 67 person by 4 occasion by 5 item design.

Separate G and D studies were conducted for each subscale and condition (placebo and medication). Across the scales and conditions, the largest portion of the variance was accounted for by person (34%-37% IO; 40%-48% OD), followed by person by occasion (21% IO, 21-25% OD), person by item (13%-15% IO, 8-12% OD), item (10%-16% IO, 1% OD), and the residual (21% IO, 18-26% OD). For this model,  $E\rho^2 = .79 - .80$  and  $\Phi = .76 - .78$  for the IO scale, and  $E\rho^2 = .82$  and  $.83$  and  $\Phi = .81 - .83$  for the IO scale.

In order to produce of range of assessment options, the sample sizes for the items and occasions facets were manipulated in a series of D studies. Notably, this criterion could not be achieved for any condition using a single-item scale, even with 20 occasions. However, adding items improved generalizability and dependability, and substantially reduced the number of occasions needed to obtain the .80 criterion, similar to the Volpe and Briesch (2012) study of multiple-item DBRs. For the IO scale, this criterion was reached for relative decisions using a 3-item scale for 7 occasions, a 4-item scale for 5-6 occasions, or 5 items for 4-5 occasions. More occasions were necessary for making absolute decisions; the criterion was obtained using a 3-item scale for 7-20 occasions, a 4-item scale for 7-10 occasions, or 5 items for 5-7 occasions. For the OD scale, this criterion was achieved using fewer occasions, with similar results for relative and absolute decisions: using a 2-item scale for 6-7 occasions, a 3-item scale for 5 occasions, or a 4 or 5-item scale for 4 occasions.

In addition to this important finding, two other findings were noteworthy. First, there appeared to be a threshold for the impact of adding items; after four or five items, generalizability and dependability gains from adding additional items leveled off. As noted by

Volpe et al. (2011), this finding contradicts the Spearman Brown prophecy, which states that reliability can be improved by lengthening the scale. Next, differences were found between generalizability and dependability of the two subscales. Volpe et al. theorized that OD represents a more homogenous construct than IO, which consists of both inattentive and hyperactivity. As noted by Gresham et al. (2010), heterogeneity may impact internal consistency. This may be supported by the obtained G study component estimates: 6%-10% of the variance was explained by items in the IO scale as compared to 1% for the OD scale.

**Summary.** Abbreviated ratings scales retain many of benefits of BRS, but require far less time to administer (Volpe et al., 2009). Although many commercially available abbreviated rating scales remain too long to be feasible for progress monitoring within an RTI context, recent research by Volpe and Gresham have focused on the development of BBRS for this purpose. This emerging research is quite promising along several lines. First, these studies indicate that scales consisting of less than 15 items may possess adequate technical features. Second, the findings from the initial G theory study by Volpe et al. (2011) indicated that accurate decisions could be made after 4 to 7 occasions. This is a notable improvement over the previously described methods, and is similar to the finding obtained with the multi-item DBRs for academic engagement in the Volpe et al. (2012) study.

Taken together, these studies suggested that brief, multiple-item scales can substantially reduce the number of assessment occasions required for accurate decision, which would improve the feasibility for progress monitoring within an RTI model. At present, this promising line of research is quite nascent and further research is required to generalize findings across other measures and with other populations. In particular, further investigation of rater effects is warranted, as the Volpe et al. study did not specify rater as a facet. Based on the results of other

G theory studies that use teachers as raters, this is likely to account for a significant portion of variance and may emerge as a potential limitation of the method.



## PURPOSE AND RATIONALE

Although several methods are currently used to assess social behavior, few are appropriate as progress monitoring tools within an RTI context (Volpe et al. 2009; 2011; Gresham et al., 2010). In reviewing the current state of social behavior assessment within problem-solving models, Chafouleas et al. (2010) noted that the field of school psychology has not reached a consensus about the best method to accomplish this goal, nor behaviors to target for this purpose. Some researchers have focused on developing general outcome measures (Chafouleas et al., 2011; Christ et al. 2009; Gresham et al.), whereas others have explored specific target behaviors and outcomes (Fabiano et al., 2009; Volpe et al., 2009; 2010; 2011). As argued by Volpe et al., both levels are necessary as part of comprehensive progress monitoring system for RTI (Volpe et al., 2011).

One promising line of research has focused on establishing single-item DBRs as general outcome measures, especially academic engagement, disruptive behavior, and compliance. However, G theory investigations of single-item DBRs have found that at between 10 to over 20 rating occasions were necessary to establish adequate reliability-like coefficients for screening or progress monitoring decisions. Yet, with multiple-item DBR scales, Volpe and Briesch were able to obtain adequate reliability-like coefficients in as few as two rating occasions. Another promising line of research has focused on developing BBRs. Although only one G theory study of this method has been conducted (Volpe et al., 2011) at present, the results of this study were similar those obtained by Volpe & Briesch (2012). Further research is necessary to replicate these encouraging findings across target behaviors and measures.

The purpose of the current study was to develop and evaluate abbreviated behavior rating scales that can potentially be used to progress-monitor social skills. As noted by Fuchs (2004),

the first step in developing a progress-monitoring tool is to establish adequate psychometric proprieties. Therefore, the primary goal of this study was to evaluate the technical adequacy of the scales with G theory. Brief behavior rating scales were to be developed for the first two domains from Social Skills Scale of the Social Skills Improvement System Rating Scales (e.g., Communication and Cooperation subscales). As our original intention was to compare the generalizability and dependability of abbreviated scales across different construction approaches, two separate methods were to be used select items from each subscale for inclusion on the BBRS: a factor analytic approach and by the recommendations of an expert panel of social skill researchers. Next, data was collected for G and D studies in a design wherein two preschool teachers rated six students in their class on all items (the 7-item Communication subscale and 6-item Cooperation subscale) on four consecutive school days. The data were analyzed in a series of G and D studies to determine the generalizability and dependability of the BBRS across variety of assessment scenarios. In particular, the number of raters, items and occasions were manipulated to determine the assessment conditions necessary to make accurate decisions.

Returning to the development of the BBRS, it was stated earlier that the individualized BBRS construction approach might be more change-sensitive than other methods (Volpe et al, 2011). This is a desirable characteristic for a progress-monitoring tool. However, Volpe also noted that it is difficult to determine the technical adequacy of individualized scales. Although the scope of this study is limited to the first two social skill domains, we recommend further development and evaluation of BBRS corresponding to the remaining Social Skill and Problem Behavior domains in future studies. Once the technical adequacy of each scale is established, the full set of BBRS should not be administered for each student, as the length would be impractical in an RTI model. Instead, after a student is identified as requiring Tier 2 interventions, the full

SSIS-RS could initially administered to determine the most problematic areas for each student. Subsequently, only the BBRS corresponding those identified domains would be administered, resulting in customized BBRS with established technical features.

In regards to establishing the technical adequacy of these scales, a strong argument was made earlier for using G theory over Classical Test Theory. To review, G theory partitions the variance in observed scores into variance due to rater, item, time, setting, method, and dimension (Cone, 1977). G theory is thus well suited for evaluating behavioral assessment as it can distinguish the individual contributions of relevant environmental variables. Moreover, D studies can be used to determine the measurement conditions necessary to make reliable norm and criterion-referenced decisions. An additional advantage of G theory is that psychometric features are not test-or sample specific; different reliability-like coefficients are calculated each time the assessment scenario is changed.

The current study evaluated the Communication and Cooperation subscales from the SSIS-RS in a series of G and D studies. The focus of the G studies was to estimate the proportion of variance associated with each component. The specified facets were raters, items (fixed facet), and occasions (p x r x i x o) in a fully-crossed mixed-model design. Although the fully-crossed design limited the number of students whose data could be analyzed, the alternative was to nest teachers and students within classrooms. This would have limited interpretation of rater-related effects (i.e., variance due to the rater facet and associated interaction components), as rater variance would be considered as error (Hoyt, 2000). As a major purpose of this study was to explore rater-related effects, nested designs were thus deemed unsuitable.

In defining the universe of admissible observations, two preschool teachers in the same classroom were selected as a sample of similar raters. Classroom teachers were selected as raters

over outside observers as teachers are the intended users of these scales. A preschool classroom was chosen as the setting for this study as these classrooms typically contain multiple teachers (e.g. lead teacher, teaching assistances, etc). The facet of occasions was included to assess the number of rating occasions needed for reliable decisions. Four occasions (each occasion was a full school day) was selected as the observed levels of this facet as this would equate to two ratings per week (over a two week period). This is similar to the recommended frequency of CBMs for Tier Two academic interventions (Hosp et al., 2007). Finally, items was specified as a fixed facet to evaluate only the extant items from Communication and Cooperation subscales from the SSIS. This is in keeping with previous research in developing BBRS by selected items from established full-length scales (Gresham et al., 2010; Volpe et al., 2009; 2010).

In developing the data collection strategy, an important consideration was the feasibility of the assessment procedures. In order to keep the time commitment required for the daily ratings reasonable (i.e., under 45 minutes per occasion) and to avoid potential fatigue effects (Harwell, 1999) it was desirable to keep the total number of items and students to a minimum (i.e., less than 15 items per student and no more than 6 students). This logic precluded the evaluation of further scales. The small sample size of students is less of a concern to generalizability theory than in experimental designs; the use of multiple facets vastly increases the number of observations (data points) per student using repeated measures design logic.

Based on the reviewed literature, it was anticipated that the largest proportion of variance would be associated with person (the objective of measurement) and the person by occasion interaction, and with smaller proportions attributed to the items and occasions facets. Previous research has found varying contributions of rater-related effects. The current study attempted mitigate these effects by specifying the raters within the same role and context (i.e., two teachers

within the same classroom) in the universe of admissible observations, as well as by increasing the rating interval to an entire day to allow adequate opportunity for raters to observe student behavior. This information was then applied to a series of D studies to determine the effect of various assessment scenarios on the dependability of the measurement system. Based on previous studies by Volpe and colleagues, it was anticipated with 5 item scales, accurate decisions could be made after three to five rating occasions.

In addition to possessing adequate psychometric properties, a progress-monitoring tool must be sensitive to detecting changes in response to an intervention. While important, this evaluation is beyond the scope of this study, as this evaluation would require a different experimental design. In G theory, the introduction of an intervention would likely confound the results (Volpe et al. 2011) whereas in a change sensitivity study, the intervention would be the independent variable. Thus, the primary goal of this study was to use G theory to establish the technical adequacy of the abbreviated scales and determine the assessment conditions required to make reliable decisions. If the decision studies suggest that reliable results can be obtained using feasible assessment procedures, further BBRS should be developed and evaluated for the remaining social skill domains on the SSIS-RS and future studies should evaluate the change sensitivity of these measures.

## METHOD

### Participants and Setting

Two teachers within one pre-kindergarten class served as the primary participants. The pre-kindergarten class was the only PreK classroom in an elementary school located in the Southeastern United States. The school served about 160 students in PreK through the fourth grade. Approximately 84.5% of students qualified for free or reduced price lunch.

Six students were selected from this classroom as secondary participants using the following procedure. First, the lead teacher completed the Social Skills Improvement System Performance Screening Guide- Preschool Version (Elliott & Gresham, 2007) for all students in the class. Students who received a score of one or two on the Prosocial Behavior item were considered as potential participants. There were four students who received a score of one and five students who received a score of two. Students with a score of one were considered first. Three students were selected as participants and fourth student was excluded due to poor attendance. Next, the students who received a score of two were considered. Of these the five students, the three with the best attendance record were selected as participants. Of the six students selected as secondary participants, three of the six students had an IEP and received services under the classification of Developmental Delay. The average age of the students was 4 years, 7 months. In accordance with IRB procedure, parental consent was obtained for all participating students prior to the study.

The setting was an inclusion pre-kindergarten classroom taught by a special education teacher and an assistant teacher with general education experience. The classroom consisted of five general education students and five students with IEPs. No specific behavioral intervention was in place for any of the students during the current study. The female lead teacher had a

bachelors in education, four years of teaching experience and was certified to teach PreK-3, and birth to age 5 as an early interventionist. The female assistant teacher had 24 years of teaching experience in a PreK setting. The assistant teacher had returned from an extended leave of absence due to medical reasons before the study started, however she had taught in this classroom for three months prior to her absence.

**Inclusion/exclusion criteria.** Two teachers from one preschool classroom served as the primary participants. The teachers were required to be present in the classroom together throughout the majority of the school day so that students were observed in the same environment at the same time. Teachers were required to spend relatively equal amount of time with students (e.g., the second teacher was not assigned to one student). Teachers who did not meet these criteria were excluded from participation.

The selected as secondary participants students were required to demonstrate adequate attendance (no more than five nonconsecutive absences during the current school year) and no planned absences during scheduled rating days. Diagnostic and/or special education status did not affect eligibility as long as the student was present in the inclusion classroom for more than 90% of the school day. This ensured that the teachers had ample opportunities to observe students. Students who received psychotropic medication or behavioral intervention were not excluded from the study provided there were no anticipated changes in treatment during the course of the study. Students who did not meet these criteria or for whom parental consent was not obtained were excluded from the study.

## **Measures**

**Communication and Cooperation subscales from the SSIS-RS Social Skill Scale.** The full 7-item Communication and 6-item Cooperation subscales (part of the Social Skills Scale)

from Social Skills Improvement System Rating Scales- Teacher Form (SSIS-RS-TF; Gresham and Elliott, 2008) was completed by both teachers for each student across each occasion. Item from these scales were reproduced on a separate sheet in order to remove items from other scales (contact author for a copy). The rating scale retained the format from the SSIS-RS. For each item, teachers indicated the relative frequency of each item as occurring never, seldom, often or almost always.

The SSIS-RS is an 83-item standardized behavior rating scale used to assess social skills and related behaviors through teacher, parent or self-report ratings. The SSIS-RS-TF consists of three scales: Social Skills (46 items), Problem Behavior (30 items) and Academic Competence (7 items). For each item on the Social Skills Scale, the teacher indicates the perceived frequency (see above) and relative importance (not important, important, and critical).

The SSIS-RS represents a major revision and restandardization of the widely used Social Skills Ratings Scale (Gresham & Elliott, 1990), with updated norms, several new scales, differentiation between skill acquisition and performance deficits, improved psychometric properties and a direct link to intervention within the Social Skills Improvement System (Elliot & Gresham, 2007; Elliott & Gresham, 2009; Gresham, Elliott, Vance, & Cook, 2011). The SSIS-RS was normed using a large standardization sample of 4,700 students aged 3 to 18, stratified to reflect the US population by race, gender and geographic region (Gresham & Elliott, 2008). As reported in the manual, the SSIS-RS has strong psychometric features, with high internal consistency (coefficient alpha of .97 for the Social Skills Scale- Teacher Form), high test-retest reliability (.84 over an average span of 42 days), good interrater reliability between pairs of teachers (.70), and offers strong evidence for convergent, discriminant and criterion-related validity with the SSRS, BASC-2, Vineland-II and other measures (see Table 2).



The SSIS-RS test manual also reports reliability estimates for each subscale. For the age 3-5 norm group ( $n = 200$ ), internal consistency was .85 for the Communication subscale and .90 for the Cooperation subscale. These values are relatively high, reflecting that items generally indicate one domain, and have low rates of random error among items. To estimate test-retest reliability, 144 students from the standardization study were rated on two occasions (mean test-retest interval of 43 days). Correlations between occasions were .85 for Communication, and .82 for Cooperation, which indicated that these behaviors are relatively stable across time. Finally, to estimate interrater reliability, ratings were obtained from pairs of teachers for 54 students from the standardization study. Interrater reliability coefficients were .63 for the Communication subscale and .60 for Cooperation.

Table 2  
Reliability Estimates as Reported in the Social Skills Improvement System Rating Scale Manual for the Teacher Form

	Internal Consistency ( $n = 200$ )	Test-Retest ( $n = 155$ )	Interrater ( $n = 54$ )
Communication Subscale	.85	.76	.60
Cooperation Subscale	.90	.85	.62
Social Skills Scale	.97	.82	.68

*Note:* These estimates are reported in the SSIS-RS technical manual (Gresham & Elliott, 2008). The correlations here reflect combined gender norms. There are five additional subscales on the Social Skill Scale that are not reported here.

**Social Skills Improvement System Performance Screening Guide.** The Social Skills Improvement System Performance Screening Guide- Preschool (SSIS-PSG; Elliott & Gresham 2007) is the classwide screening component within the Social Skills Improvement System. In this criterion-referenced measure, teachers rate each student in four areas related to social skills: Prosocial Behavior, Motivation to Learn, Early Reading Skills and Early Math Skills. Each area is rated on a four-point scale with specific descriptions of skills and behaviors described at each

of the four levels, with 1 representing the lowest level of skills, and 4 representing the highest level. A score of 1 denotes a high level of concern, requiring additional intervention. A score of 2 indicates moderate levels of concern and some additional instruction. Finally, scores of 3 or 4 indicate adequate or superior functioning; no additional instruction is necessary. The SSIS-PSG was field-tested using a subsample of 138 teachers from the SSIS-RS standardization sample. According to the SSIS-RS manual (2008), the SSIS-PGS has high test-retest reliability, with intraclass correlations coefficients for the Elementary level ranging from .69 to .74, and moderate inter-rater reliability with intraclass correlations coefficients ranging from .55 to .58.

### **Procedures**

**Participant identification.** The primary researcher contacted elementary school principals in the Boston metropolitan area and South Louisiana region to identify potential preschool classrooms for data collection. These regions were chosen for convenience (i.e. proximity to the researcher and her colleagues). After a school was identified and administrative approval was obtained, the principal recommended a pair of preschool teachers as potential participants. In this school, there was only one preschool classroom. The teachers were provided with a written overview of the study. A research assistant (a graduate student in school psychology) met with these teachers to discuss the general purpose and parameters of the study, and answer any questions about the study. At this time, informed consent was obtained from the teachers. Afterwards, the lead teacher completed the Social Skills Improvement System Performance Screening Guide (Elliott & Gresham, 2007) in order to identify students as potential participants, as detailed previously. Figure 3 below displays a flow chart of the selection process.

**Development of abbreviated rating scales.** The intended goal of this phase was to develop two sets of abbreviated, 5-item scales using two different methods to select items for

inclusion (factor analytic and expert panel ratings construction methods). It was originally proposed to develop the abbreviated scales first, and then conduct a separate set of G and D studies on each method to determine which method produced adequate generalizability coefficients using the most efficient assessment procedure (e.g. the fewest number of items and occasions). Upon further deliberation, the procedure was altered so that the full 7-item Communication scale and 6-item Cooperation scale were administered to teachers, and the full model G and D studies were conducted using the full-item scales. This maximized the information that could be obtained. The previous decision to construct 5-item abbreviated scale used an arbitrary number of items. Instead, information from D studies with the full-item scales could help determine the optimal number of items for inclusion on the abbreviated scales.

**Data collection.** The teachers independently completed the Communication and Cooperation subscales rating form for each student on four consecutive school days. As recommended by the SSIS-RS manual, both teachers were familiar with the participants. Teachers were instructed to review a copy of the rating scale on the morning of each rating occasion, and observe whether the participants performed these items throughout the school day. To preserve the authenticity of a typical rating scenario, teachers were asked to observe the students while conducting normal teaching duties. To ensure that the ratings were completed independently, teachers reported to the principal at the end of the day and completed the rating forms in the office area under the supervision of the principal. Ratings were conducted in the same order each day to keep control for rater fatigue effects. Teachers were instructed to complete the ratings based on their observations for that day only.

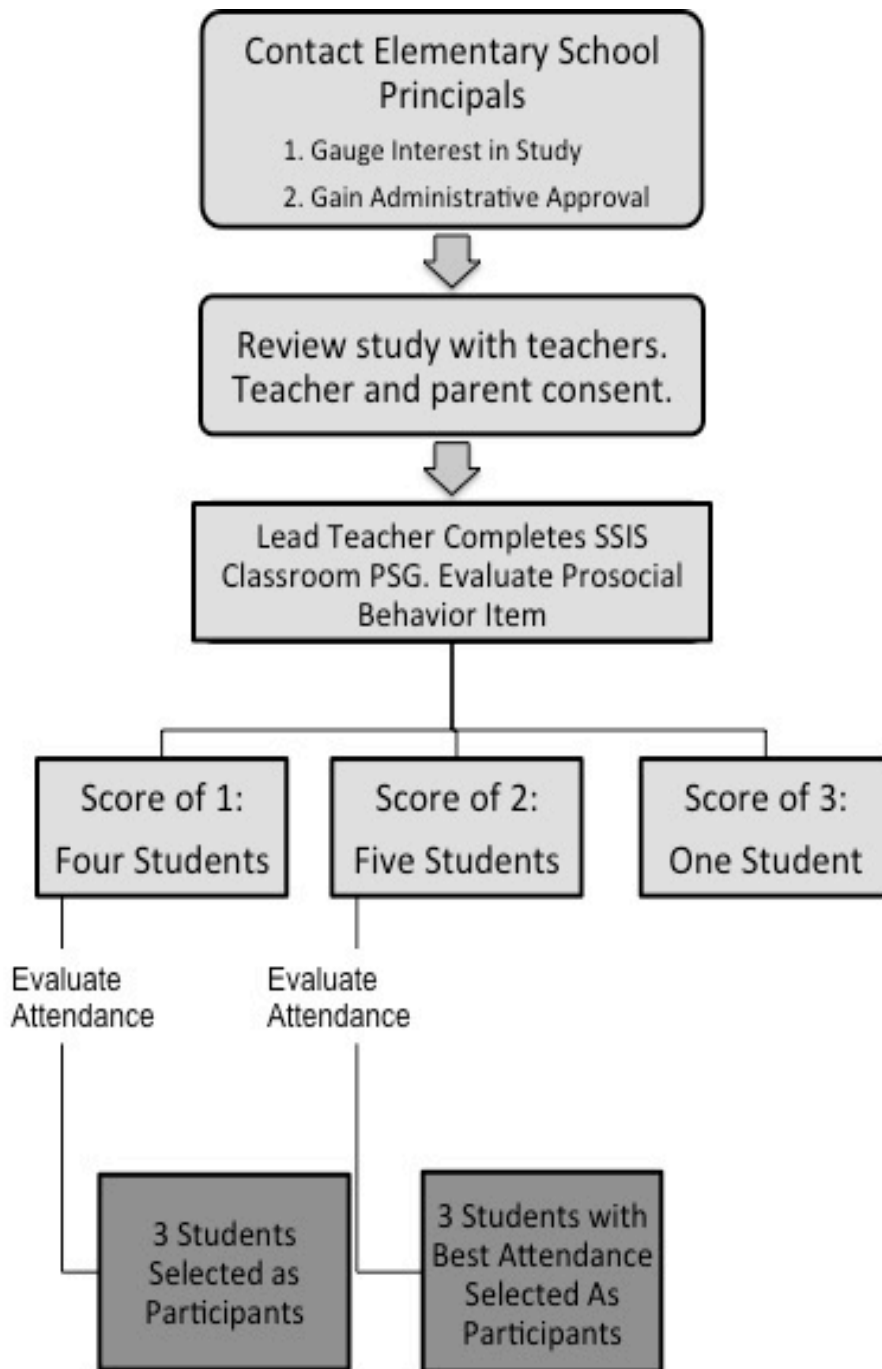


Figure 3. Participant Selection Flowchart

If one student was absent on a given day, teachers were instructed to complete ratings for the other students and rate the absent student on the first day the student returned to school. The next rating occasion could not be completed until all students were rated for that occasion. If

more than one student was absent, teachers were instructed to wait until the at least five of the six students were present before completing a rating for that occasion. Completed ratings were securely stored in the principal's office and were collected by the research assistant.

Upon completion of data collection phase, teachers received a \$75 gift certificate as compensation for their time.

## STATISTICAL PROCEDURES AND ANALYSES

### Generalizability Studies

Separate G studies were conducted for the full-item Communication and Cooperation subscales using fully-crossed person (6 students) by rater (2 teachers) by item (7 for Communication, 6 for Cooperation) by occasion (4 consecutive days) mixed model design. Person was designated as the object of measurement, raters and occasions modeled as random components, and items modeled as a fixed component. Hereafter, this design will be referred to as the full model G study. All calculations were performed in EduG (Swiss Society for Research in Education Working Group, 2010).

Estimated variance components for full model G study were calculated for the following facets and interactions: person, rater, item, occasion, person x rater, person x item, person x occasion, rater x item, rater x occasion, item x occasion, person x rater x item, person x item x occasion, rater x item x occasion, and the residual four-way person x rater x item x occasion plus error interaction (see Table 3).

Based on the results of the full model G study, a second series of G studies were conducted for each subscale to separately examine the sources of variance for each teacher. When examining results by teacher, this reduced the original study design to a fully crossed person (6 students) by item (7 for Communication, 6 for Cooperation) by occasion (4 consecutive days) mixed model design. Hereafter, this design is referred to as the reduced model G study. Estimated variance components were calculated for the following facets and interactions: person, item, occasion, person x item, person x occasion, item x occasion, and the residual three-way person x rater x item x occasion plus error interaction.

Table 3  
Sources of Variation in p X r X i X o Fully Crossed Mixed-Model Design with Items Fixed

Facet	Type of variation	Variance notation
Persons	Object of measurement	$\sigma_p^2$
Raters ( <i>r</i> )	Variability due to inconsistency between raters	$\sigma_r^2$
Items ( <i>i</i> )	Variability due to inconsistency in item difficulty	$\sigma_i^2$
Occasions ( <i>o</i> )	Variability due to inconsistency across occasions	$\sigma_o^2$
<i>p X r</i>	Variability due to p x r interaction	$\sigma_{pr}^2$
<i>p X i</i>	Variability due to p x i interaction	$\sigma_{pi}^2$
<i>p X o</i>	Variability due to p x o interaction	$\sigma_{po}^2$
<i>r X i</i>	Variability due to r x i interaction	$\sigma_{ri}^2$
<i>r X o</i>	Variability due to r x o interaction	$\sigma_{ro}^2$
<i>i X o</i>	Variability due to i x o interaction	$\sigma_{io}^2$
<i>p X r X i</i>	Variability due to p x r x i interaction	$\sigma_{pri}^2$
<i>p X r X o</i>	Variability due to p x r x o interaction	$\sigma_{pro}^2$
<i>p X i X o</i>	Variability due to p x i x o interaction	$\sigma_{pio}^2$
<i>r X o X i</i>	Variability due to r x o x i interaction	$\sigma_{rio}^2$
<i>p X r X i X o + e</i>	Variability due to r x o x i interaction plus error	$\sigma_{prio,e}^2$

**Statistical procedures.** The estimated variance components for the G studies were derived using ANOVA sum of squares in EduG (Swiss Society for Research in Education Working Group, 2010), a freeware statistical program developed specifically for G theory analyses. This analysis provides the associated mean squares and the degrees of freedom used to estimate the variance for each component using the rules as described by Brennan (2010) in the second chapter of this proposal. For fixed facets, EduG employs the “averaging over conditions of the fixed facets” approach described by Shavelson and Webb (1991, p. 67-70) to calculate the variance components for mixed model designs. Tables 4 and 5 displays the specific formulas used to derive estimates for each component in the current study.

## Decision Studies

Using results obtained from the G studies, a series of D studies were conducted for each subscale (full model D studies). To review, the goal of a D study is to define the universe of generalization for a specified purpose and then analyze how changing various facets of measurement impacts decisions made from the resulting data. As previously discussed, we decided to capitalize the use of D studies to inform the selection of an optimal number of items for inclusion on the abbreviated scale by conducting the analysis using the full-item scales.

The universe of generalization for the full model D studies shared the same person by rater by item by occasion design as the G study.  $N_r$  represented a random sample from a universe of all preschool teachers from similar educational backgrounds, teaching at similar schools. The seven items on the Communication subscale and the six items on the Cooperation subscale represented the entire universe of acceptable items for each scale. This universe was modeled as a fixed facet as we were primarily interested in evaluating the extant items from the SSIS-RS scale. Finally, occasion was modeled as a random facet.  $N_o$  represents the universe consisting of any typical school day.

For the full model D studies, facets were manipulated first separately and then simultaneously. The items facet was manipulated to explore the minimal amount of items necessary for inclusion on an abbreviated scale while maintaining obtain acceptable (e.g.,  $\geq .80$ ) generalizability ( $E\rho^2$  for relative decisions) and dependability ( $\Phi$  for absolute decisions) coefficients. The occasions facet was manipulated to determine how many measurement occasions optimized generalizability and dependability.

The rater facet was manipulated to see if one rater could produce reliable results, and if not, how many raters would be required to achieve this outcome.



Table 4  
G-Study Variance Components Using ANOVA with Type III Sums of Squares

Facet	SS	Degrees of Freedom	MS	Estimated Variance Component
Persons ( $p$ )	$SS_p$	$n_p - 1$	$SS_p/df_p$	$\hat{\sigma}^2(p) = \frac{MS(p) - MS(pr) - MS(pi) - MS(po) + MS(pri) + MS(pro) + MS(pio) - MS(prio)}{n_r n_i n_o}$
Raters ( $r$ )	$SS_r$	$n_r - 1$	$SS_r/df_r$	$\hat{\sigma}^2(r) = \frac{MS(r) - MS(pr) - MS(ri) - MS(ro) + MS(pri) + MS(pro) + MS(rio) - MS(prio)}{n_p n_i n_o}$
Items ( $i$ )	$SS_i$	$n_i - 1$	$SS_i/df_i$	$\hat{\sigma}^2(i) = \frac{MS(i) - MS(pi) - MS(ri) - MS(io) + MS(pri) + MS(pio) + MS(rio) - MS(prio)}{n_p n_r n_o}$
Occasions ( $o$ )	$SS_o$	$n_o - 1$	$SS_o/df_o$	$\hat{\sigma}^2(o) = \frac{MS(o) - MS(po) - MS(ro) - MS(io) + MS(pro) + MS(pio) + MS(rio) - MS(prio)}{n_p n_r n_i}$
$p \times r$	$SS_{pr}$	$(n_p - 1)(n_r - 1)$	$SS_{pr}/df_{pr}$	$\hat{\sigma}^2(pr) = \frac{MS(pr) - MS(pri) - MS(pro) + MS(prio)}{n_i n_o}$
$p \times i$	$SS_{pi}$	$(n_p - 1)(n_i - 1)$	$SS_{pi}/df_{pi}$	$\hat{\sigma}^2(pi) = \frac{MS(pi) - MS(pri) - MS(pio) + MS(prio)}{n_r n_o}$
$p \times o$	$SS_{po}$	$(n_p - 1)(n_o - 1)$	$SS_{po}/df_{po}$	$\hat{\sigma}^2(po) = \frac{MS(po) - MS(pro) - MS(pio) + MS(prio)}{n_r n_i}$
$r \times i$	$SS_{ri}$	$(n_r - 1)(n_i - 1)$	$SS_{ri}/df_{ri}$	$\hat{\sigma}^2(ri) = \frac{MS(ri) - MS(pri) - MS(rio) + MS(prio)}{n_p n_o}$
$r \times o$	$SS_{ro}$	$(n_r - 1)(n_o - 1)$	$SS_{ro}/df_{ro}$	$\hat{\sigma}^2(ro) = \frac{MS(ro) - MS(pro) - MS(rio) + MS(prio)}{n_p n_i}$
$i \times o$	$SS_{io}$	$(n_i - 1)(n_o - 1)$	$SS_{io}/df_{io}$	$\hat{\sigma}^2(io) = \frac{MS(io) - MS(pio) - MS(rio) + MS(prio)}{n_p n_r}$

Table 4 Continued

Facet	SS	Degrees of Freedom	MS	Estimated Variance Component
$pXrXi$	$SS_{pri}$	$(n_p - 1)(n_r - 1)(n_i - 1)$	$SS_{pri} / df_{pri}$	$\hat{\sigma}^2(pri) = \frac{MS(pri) - MS(prio)}{n_o}$
$pXrXo$	$SS_{pro}$	$(n_p - 1)(n_r - 1)(n_o - 1)$	$SS_{pro} / df_{pro}$	$\hat{\sigma}^2(pro) = \frac{MS(pro) - MS(prio)}{n_i}$
$pXiXo$	$SS_{pio}$	$(n_p - 1)(n_i - 1)(n_o - 1)$	$SS_{pio} / df_{pio}$	$\hat{\sigma}^2(pio) = \frac{MS(pio) - MS(prio)}{n_r}$
$rXoXi$	$SS_{rio}$	$(n_r - 1)(n_o - 1)(n_i - 1)$	$SS_{rio} / df_{rio}$	$\hat{\sigma}^2(rio) = \frac{MS(rio) - MS(prio)}{n_p}$
$pXrXiXo,e$	$SS_{prio}$	$(n_p - 1)(n_r - 1)(n_i - 1)(n_o - 1)$	$SS_{prio} / df_{prio}$	$\hat{\sigma}^2(prio) = MS(prio)$

Table 5

## Variance Notation for G and D Studies, Averaging Across Fixed Facet (Items) Method

p X r X i X o completely crossed random design	p X r X i X o completely crossed with i fixed; r, o random	p X R X I X O completely crossed with I fixed; R, O random		
Facet	Variance Notation averaged over i	Estimated variance components averaging over items	Estimated Variance Components averaging across items	Contributions to universe score and error variances
Persons (p)	$\sigma_{p^*}^2$	$\sigma_p^2 + \frac{1}{n_i} \sigma_{pi}^2$	$\sigma_p^2 + \frac{1}{n_i} \sigma_{pi}^2$	$\Gamma$
Raters (r)	$\sigma_{r^*}^2$	$\sigma_r^2 + \frac{1}{n_i} \sigma_{ri}^2$	$\frac{\sigma_r^2 + \frac{1}{n_i} \sigma_{ri}^2}{n_r}$	$\Delta$
Items (i)				
Occasions (o)	$\sigma_{o^*}^2$	$\sigma_o^2 + \frac{1}{n_i} \sigma_{io}^2$	$\frac{\sigma_o^2 + \frac{1}{n_i} \sigma_{io}^2}{n_o}$	$\Delta$
p X r	$\sigma_{pr^*}^2$	$\sigma_{pr}^2 + \frac{1}{n_i} \sigma_{pri}^2$	$\frac{\sigma_{pr}^2 + \frac{1}{n_i} \sigma_{pri}^2}{n_r}$	$\Delta \delta$
p X i				$\Gamma$
p X o	$\sigma_{po^*}^2$	$\sigma_{po}^2 + \frac{1}{n_i} \sigma_{poi}^2$	$\frac{\sigma_{po}^2 + \frac{1}{n_i} \sigma_{poi}^2}{n_o}$	$\Delta \delta$
r X i				$\Delta$
r X o	$\sigma_{ro^*}^2$	$\sigma_{ro}^2 + \frac{1}{n_i} \sigma_{roi}^2$	$\frac{\sigma_{ro}^2 + \frac{1}{n_i} \sigma_{roi}^2}{n_r n_o}$	$\Delta$

Table 5 Continued

Facet	Variance Notation averaged over i	Estimated variance components averaging over items	Estimated Variance Components averaging across items	Contributions to universe score and error variances
p X r X i X o completely crossed random design	p X r X i X o completely crossed with i fixed; r, o random	p X R X I X O completely crossed with I fixed; R, O random		
p X r X i				$\Delta \delta$
p X r X o	$\sigma_{pro,e}^2$	$\sigma_{pro}^2 + \frac{1}{n_i} \sigma_{prio,e}^2$	$\frac{\sigma_{pro}^2 + \frac{1}{n_i} \sigma_{prio,e}^2}{n_r n_o}$	$\Delta \delta$
p X i X o				$\Delta \delta$
r X o X i				$\Delta$
p X r X i X o + e				$\Delta \delta$

Note:  $\Gamma$  contributes to universe score variance,  $\Delta$  to relative error variance, and  $\delta$  to absolute error variance

After results were obtained, another series of D studies addressed all three facets together to determine the full set of assessment conditions to optimize generalizability and dependability.

After consideration of the previous results, a second series of D studies were conducted for each teacher using an identical design as the reduced G model studies (i.e., person by item by occasion mixed model design with items as a fixed facet and occasions as a random facet). Hereafter, this design is referred to as the reduced model D study. In the reduced model D studies, the facets of items and occasions were systematically manipulated. However, interpreting the outcomes within rater restricts the universe of generalization to the universe of ratings completed by that individual, on the extant items, across any typical school day. The implications of this change will be reviewed in detail in the discussion section.

The purpose of the reduced model D studies was to determine the assessment conditions necessary to achieve adequate generalizability using ratings completed by one individual rater. This outcome is particularly important as ratings are often completed by one teacher within the context of natural classroom environments (Chafouleas et al., 2007). We were interested to examine whether the individual teachers could make accurate decisions. The  $\geq .80$  threshold was employed for adequate generalizability and dependability coefficients, as this is commonly used as the criterion for making reliable decisions for screening and progress monitoring purposes (Saliva, et al., 2004).

**Statistical procedures.** Using the same completely crossed, mixed design as the G studies, the generalizability and dependability coefficients were calculated for each scale. First, D study variance component estimates were converted from the G study estimates by dividing each components by the product of D study sample sizes ( $n'$ ) for each index that makes up that component, excluding person (see Table 4). Next, the universe score variance =  $\sigma^2(\tau)$ , absolute

error variance =  $\sigma^2(\Delta)$  , and relative error variance =  $\sigma^2(\delta)$  were calculated using Shavelson and Webb's (1991) averaging across fixed facets approach. Applying the mixed model D study variance components to the formulas for obtaining universe score, absolute and relative error variances yielded the following:

$$\sigma^2(\tau) = \sigma_p^2 + \frac{1}{n_i} \sigma_{pi}^2 \quad (13)$$

$$\begin{aligned} \sigma^2(\Delta) = & \frac{\sigma_r^2 + \frac{1}{n_i} \sigma_{ri}^2}{n_r} + \frac{\sigma_o^2 + \frac{1}{n_i} \sigma_{io}^2}{n_o} + \frac{\sigma_{pr}^2 + \frac{1}{n_i} \sigma_{pri}^2}{n_r} \\ & + \frac{\sigma_{po}^2 + \frac{1}{n_i} \sigma_{pio}^2}{n_o} + \frac{\sigma_{ro}^2 + \frac{1}{n_i} \sigma_{rio}^2}{n_r n_o} + \frac{\sigma_{pro}^2 + \frac{1}{n_i} \sigma_{prio,e}^2}{n_r n_o} \end{aligned} \quad (14)$$

$$\sigma^2(\Delta) = \frac{\sigma_{pr}^2 + \frac{1}{n_i} \sigma_{pri}^2}{n_r} + \frac{\sigma_{po}^2 + \frac{1}{n_i} \sigma_{pio}^2}{n_o} + \frac{\sigma_{pro}^2 + \frac{1}{n_i} \sigma_{prio,e}^2}{n_r n_o} \quad (15)$$

As noted in the above equations,  $\sigma^2(\tau)$  consisted of the variance components for person and the interaction between persons and the fixed item facet. Similarly,  $\sigma^2(\Delta)$  was composed of all components besides those contained in  $\sigma^2(\tau)$ , and finally,  $\sigma^2(\delta)$  consisted of all interactions between persons and other facets. Figure 2 displays Venn diagrams of the sources of variance. After obtaining the variances for the universe score, absolute error and relative error, the generalizability and dependability coefficients were obtained using the formulas detailed in equations 9 and 10. Finally, these steps were repeated for each D study described above.

## RESULTS

One hundred percent ( $n = 624$ ) of the 624 possible data points (6 students x 2 raters x 13 items x 4 occasions) were collected. On the fourth (and final) occasion, one student was absent. Following the procedure for student absence, this student was rated on the following day. This procedure affected 26, or 4.17% of the total data points. The overall means and standard deviations were  $M = 1.35$  ( $SD = .43$ ) for the Communication subscale and  $M = 1.43$  ( $SD = .44$ ) for the Cooperation Subscales. Mean ratings for the individual items on each occasion are displayed in Table 6.

### Generalizability Theory Analyses

**Full model G studies.** The full data set was explored by conducting G and D studies for all items on the 7-item Communication and 6-item Cooperation subscale separately. The full model reflects the observed study conditions in which all six students ( $p$ ), were independently rated by both teachers ( $r$ ) on all items ( $i$ ) on all four days ( $o$ ). Raters and occasions were modeled as random facets and items as a fixed facet. The full model G studies results are summarized in Table 7.

For the Communication subscale, more than 45% of the total variance was attributed to rater-related effects. In particular, the rater by item interaction (15.7%) and the person by rater interaction (12.3%) represented the first and third largest contributors to the explained variance. These results indicated some moderate differences in how teachers ranked items and students. The main effect of rater (across students, occasions and items) was 5.5%, suggesting some minor difference in how the two teacher rated the overall level of student behavior.

In contrast, variance due to individual differences between students was lower than anticipated, with person (the object of measurement) accounting for 13.7% of the total variance,

Table 6  
Descriptive Statistics For Items Across Occasions

Item	Occasion 1		Occasion 2		Occasion 3		Occasion 4	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Communication								
Says please	1.75	0.62	1.67	0.65	1.42	0.51	1.33	0.65
Responds to others	1.58	0.51	1.58	0.67	1.50	0.52	1.58	0.90
Appropriate tone	1.58	0.79	1.67	0.78	1.42	0.51	1.50	0.52
Takes turns convo	1.17	0.39	1.42	0.51	1.42	0.67	1.17	0.39
Says thank you	1.67	0.65	1.67	0.65	1.50	0.52	1.58	0.79
Eye contact talking	1.67	0.49	1.67	0.49	1.75	0.87	1.67	0.65
Gestures/body language	1.83	0.83	1.83	0.83	2.08	0.90	1.83	0.83
Cooperation								
Follows directions	1.92	0.67	1.92	0.67	2.00	0.74	1.58	0.67
Completes tasks	1.92	0.67	1.58	0.79	1.75	0.75	1.83	0.83
Participates appropriately	2.33	0.78	2.17	0.72	2.33	0.49	2.08	0.90
Ignores distractions	1.75	0.62	1.58	0.67	1.92	0.67	1.58	0.67
Pay attention	1.67	0.49	1.00	0.43	1.17	0.58	0.92	0.29
Follows rules	1.92	0.67	1.67	0.65	1.92	0.79	1.58	0.67



Table 7  
Full Model G Study Variance Component Estimates

Component	SSIS-RS Social Skill Subscale			
	Communication		Cooperation	
	Var	% Var	Var	%Var
Person	0.08	(13.7%)	0.15	(24.0%)
Rater	<b>0.03</b>	<b>(5.5%)</b>	<b>0.06</b>	<b>(9.2%)</b>
Item	0.00	(0%)	0.08	(11.8%)
Occasion	0.00	(0%)	0.01	(2.3%)
Person x rater	<b>0.07</b>	<b>(12.3%)</b>	<b>0.05</b>	<b>(8.4%)</b>
Person x item	0.01	(2.0%)	0.02	(2.8%)
Person x occasion	0.04	(7.2%)	0.01	(1.8%)
Rater x item	<b>0.09</b>	<b>(15.7%)</b>	<b>0.03</b>	<b>(4.7%)</b>
Rater x occasion	<b>0.00</b>	<b>(.1%)</b>	<b>0.00</b>	<b>(0.0%)</b>
Item x occasion	0.00	(0%)	0.00	(0.0%)
Person x rater x item	<b>0.02</b>	<b>(3.9%)</b>	<b>0.00</b>	<b>(0.0%)</b>
Person x rater x occasion	<b>0.03</b>	<b>(4.7%)</b>	<b>0.08</b>	<b>(12.6%)</b>
Person x item x occasion	0.00	(0.0%)	0.02	(2.5%)
Rater x item x occasion	<b>0.02</b>	<b>(3.7%)</b>	<b>0.02</b>	<b>(3.1%)</b>
Person x rater x item x occasion + e	0.18	(31.1%)	0.11	(16.8%)
$E\rho^2$	.62		.79	
$\Phi$	.55		.68	

Note. Var = variance component using ANOVA sum of squares, % Var = percentage of total variance. Variance components are rounded to nearest hundredth place. Variance components that include rater effects are in boldface, with the exception of person x rater x item x occasion as this component includes unspecified error.

and the person by occasion interaction accounted for an additional 7.2%. These results indicated some moderate differences in students' communication abilities, as well as some change in the relative standing of students across occasions. The remaining facets and interactions accounted for less than 5% of the variance. Notably, item, occasion and interactions between these facets were negligible, which indicated overall consistency of communicative behavior across items and occasions. Finally, the residual error term accounted for 31.1% of the total variance, which included variance from the four-way person by rater by item by occasion interaction as well as unexplained variance. For the full scale model, the generalizability and dependability

coefficients obtained with enacted data collection procedure were  $E\rho^2 = .62$  and  $\Phi = .55$  respectively. The obtained SEMs were  $\pm .22$  for relative decisions and  $\pm .25$  for absolute decisions.

For the Cooperation scale, the person facet accounted for the largest proportion of the total variance (24%). This was desirable, as the combined person and person by item interaction (2.8%) represented universe score variance, analogous to true score variance in CTT. However, the rater-related effects were substantial, with the rater facet accounting for 9.2% of the total variance and another 8.4% by the person by rater interaction. This suggests some overall differences in rater leniency, as well as differences in the rank ordering of students between teachers. The three-way person by rater by occasion interaction explained an additional 12.6% of the variance. In contrast, the person by occasion interaction explained only 1.8% of the variance. In other words, some variance was explained by differences in how the two teachers ranked ordered student cooperation on different days, whereas the overall ranking of students (collapsed across raters) was fairly consistent across occasions.

Unlike the Communication scale, the items facet had a sizeable impact, (11.8%) and the rater by item interaction was lower (4.7%). This indicates that there were overall differences in the item ratings (the overall frequency rating for each item, collapsed across the students and other facets), but only minor differences in how the two teachers ranked the overall frequency of the items. Finally, the residual accounted for 16.8% of the variance. For the full scale model, the generalizability and dependability coefficients obtained with enacted procedure were  $E\rho^2 = .79$  and  $\Phi = .68$  respectively. The obtained SEMs were  $\pm .20$  for relative decisions and  $\pm .27$  for absolute decisions.

To better understand rater-related effects, the average ratings for each teacher were graphed and visually compared (see Figure 4). First, the lead teacher generally rated the students as performing communication and cooperation behaviors more frequently than the assistant teacher. Next, when examining the average scores for each student (across items and occasions), the two teachers agreed on the frequency of behaviors for students one, two and three on both scales, and for student six on the Communication scale, but were discrepant for students four and five on both scales. Likewise, the teachers had similar ratings for items 1, 3, 4 and 5 on the Communication scale and items 4, 5, and 6 on the Cooperation subscale, but differed on the remaining items. Finally, as indicated on the graphs on right side of Figure 4, the teachers rated the overall level of behavior fairly consistently across occasions (averaged over students and items) on both scales.

**Full-model D studies.** Before constructing the abbreviated scales, it was important to first assess if adequate generalizability and dependability ( $\geq .80$ ) could be obtained for the full item subscales using feasible assessment procedures. Therefore, a series of full model D studies were conducted wherein each facet was systematically manipulated. First, the number of raters, items, and occasions were each manipulated separately, while holding the other facets constant at the observed levels (e.g., two raters, six or seven items, and/or four occasions). This allowed for exploration of each facet's contribution to generalizability and dependability. Based on the results of the full model G studies (i.e., the large proportion of total variance attributed to rater related effects), it was anticipated that increasing the number of raters would produce the biggest impact. Results of these manipulations are displayed in Figure 5.

As anticipated, adding additional raters produced the largest gains in the generalizability and dependability coefficients for both subscales. However, the extent of this increase relative to

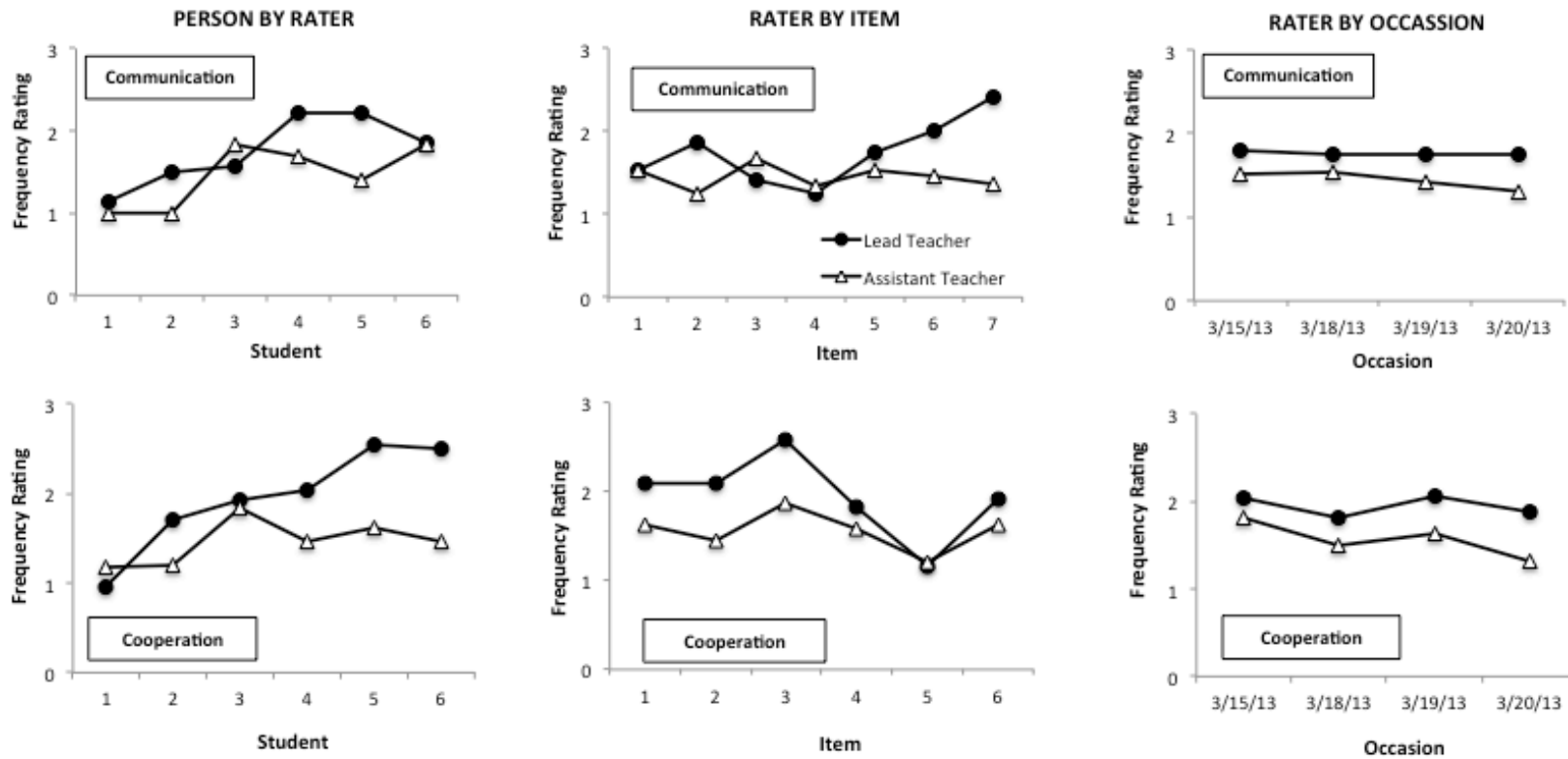


Figure 4. Mean Frequency Ratings by Rater for the Full 7-Item Communication and 6-Item Cooperation Subscales. Rater averages were calculated from the obtained data with each teacher rating the same 6 students on all 13 items on all four occasions. Frequency ratings reflect how often a behavior is performed: a score of 0 indicates never, 1 is seldom, 2 is often and 3 almost always. The graphs on the left depict the average score for each student, collapsed across item and occasion. The middle graphs shows the average score for each item, collapsed across student and occasions. For description of each item, refer to Appendix A. The graphs on the right displays the average score for each occasion, collapsed across students and items.

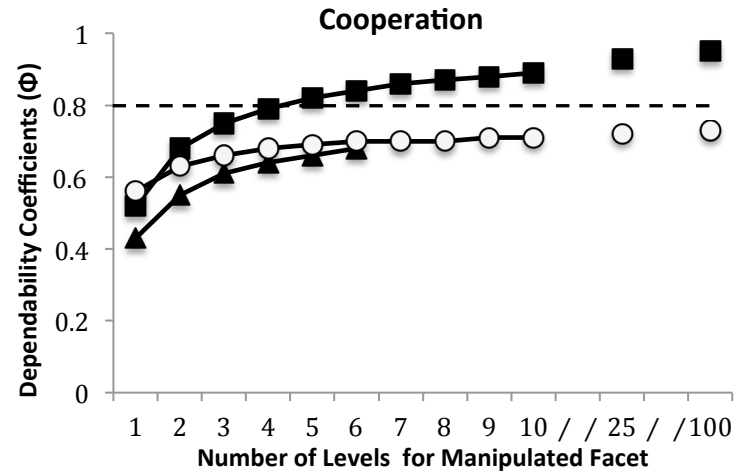
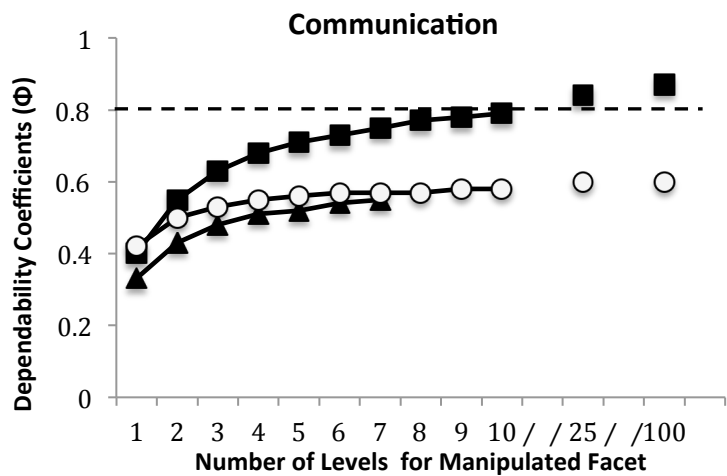
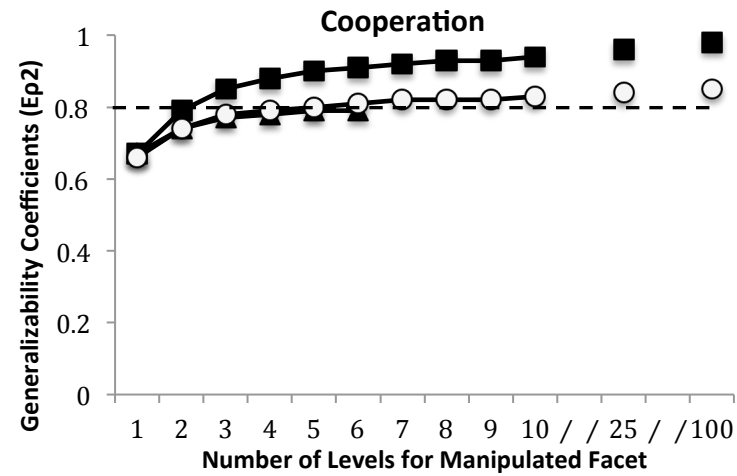
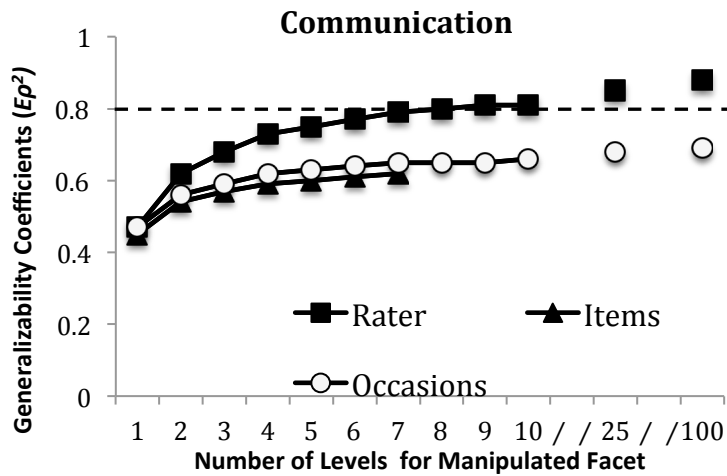


Figure 5. Full Model Dependability Studies. Each facet was manipulated independently, holding other facets constant at observed levels (e.g., 2 raters, 7 items for Communication; 6 for Cooperation, 4 occasions). The dashed lines indicate the  $\geq .80$  criterion for adequate generalizability and dependability.

other facets was surprising. For the Communication scale, eight raters were required to achieve adequate generalizability and twelve raters were required for adequate dependability.

Comparatively, increasing the number of occasions produced only marginal gains; even with 100 occasions (and two raters and all seven items), generalizability and dependability remained well below the criterion ( $E\rho^2 = .69$ ;  $\Phi = .60$ ). For the Cooperation scale, three raters were required to achieve adequate generalizability and five for adequate dependability. Adequate generalizability was also achieved after eight occasions (using two raters and all six items), but adequate dependability was never achieved by adding additional occasions, even with 100 occasions. Finally, reductions to the number of items were detrimental to generalizability and dependability coefficients, particularly when the scale consisted of three or fewer items.

Next, a series of follow-up D studies was conducted in which the facets were simultaneously manipulated to see if adequate generalizability and dependability could be obtained with fewer raters. Starting with the amount of raters reported above, the number of items (from 1 to the full 6 or 7-item scale) and occasions (1-10 occasions) were systematically manipulated. This process was repeated, removing one rater each time until the obtained generalizability and dependability coefficients fell below the .80 criterion.

This process resulted in a range of assessment conditions in which the .80 criterion was achieved. For the Communication scale, adequate generalizability coefficients were obtained with 5 raters, 6 items and 9 occasions. Additional options included 2 items, 8 raters and 10 occasions; 5 occasions, 7 raters, and 6 items; or with 5 occasions, 4 items and 8 raters. Adequate dependability required additional levels of the facets: 7 raters, 6 items, and 10 occasions; 3 items, 10 raters, and 10 occasions; 3 items, 11 raters and 8 occasions; or 4 occasions, 12 raters and 6 items.

For the Cooperation scale, adequate generalizability could be feasibly obtained with 2 raters and several combinations of items and occasions (3 items and 8 occasions; 8 occasions and 4 items; or with 5 items and 5 occasions). Additional options include 3 raters, 2 items and 5 occasions, or 2 occasions, 5 items and 3 raters. Adequate dependability required additional raters and produced fewer options: 4 raters with 5 items and 7 occasions, or with 6 items and 5 occasions.

**Reduced model G studies.** A second set of G studies was conducted to compare and contrast sources of variability for each teacher for the Communication and Cooperation subscales. Results from these studies are displayed in Tables 8 and 9.

Table 8  
Reduced Model G Study Variance Component Estimates for Communication

Component	SSIS-RS Communication Subscale			
	Lead Teacher		Assistant Teacher	
	Var	% Var	Var	% Var
Person	0.17	(24.9%)	0.13	(39.9%)
Item	0.11	(17.1%)	0.01	(2.9%)
Occasion	0.00	(0.0%)	0.00	(0.0%)
Person x item	0.03	(4.8%)	0.04	(10.8%)
Person x occasion	0.06	(9.5%)	0.07	(22.1%)
Item x occasion	0.02	(3.3%)	0.00	(0.4%)
Person x item x occasion + e	0.27	(40.3%)	0.08	(23.9%)
$E\rho^2$		.91		.88
$\Phi$		.91		.88

*Note.* Var = variance component using ANOVA sum of squares, % Var = percentage of total variance. Variance components are rounded to nearest hundredth place.

For the Communication scale, the within rater interpretation led to a sizeable improvement in universe score variance particularly for the assistant teacher (13.7% for the full model, as compared with 24.9% for the lead teacher and 39.9% for the assistant teacher). The residual error term increased to 40.3% for the lead teacher, but decreased to 23.9% for the

assistant teacher. For the lead teacher, a larger proportion of the variance was explained by overall differences (across students and occasions) between items than by differences in the rank ordering of items between students or than by changes in student ranking across occasions (17.1% for item, 4.8% for person by item, and 9.5% for person by occasion). For the assistant teacher, the person by occasion interaction explained more variance (22.1%) than the items facet and the person by item interaction combined (2.9% and 10.8% respectively). Given the enacted model with 6 students, 7 items and 4 occasions, both teachers could make accurate screening (SEM = .13 for the lead teacher; = .88, SEM = .14 for the assistant teacher) and progress monitoring decisions ( $\Phi = .91$ , SEM = .13 for the lead teacher;  $\Phi = .88$ , SEM = .14 for the assistant teacher).

Table 9  
Reduced Model G Study Variance Component Estimates for Cooperation

Component	SSIS-RS Cooperation Subscale			
	Lead Teacher		Assistant Teacher	
	Var	% Var	Var	% Var
Person	0.31	(43.4%)	0.10	(24.1%)
Item	0.17	(22.9%)	0.03	(8.2%)
Occasion	0.00	(0.2%)	0.02	(6.0%)
Person x item	0.01	(.8%)	0.02	(5.9%)
Person x occasion	0.06	(8.5%)	0.12	(30.2%)
Item x occasion	0.03	(4.6%)	0.00	(0.1%)
Person x item x occasion + e	0.14	(19.6%)	0.08	(25.5%)
$E\rho^2$	.95		.76	
$\Phi$	.95		.73	

*Note.* Var = variance component using ANOVA sum of squares, % Var = percentage of total variance. Variance components and percentages are rounded to nearest hundredth place.

For the Cooperation scale, the within rater interpretations led to a sizable improvement in the universe score variance for the lead teacher, but not for the assistant teacher (24% full model; 43.4% lead teacher; 24.1% assistant teacher). Furthermore, the variance due to the residual error only increased slightly for each rater over the full scale G study (16.8% full model; 19.6% lead



teacher; 25.5% assistant teacher). The proportion of variance due to items facet and the person by item interaction was nearly identical to the Communication scale (22.9% and .8% respectively for the lead teacher; 8.2% and 5.9% for the assistant teacher). For the lead teacher, the person by occasion interaction (i.e., different students were ranked as most cooperative on different days) explained a similar proportion of variance (8.5%) to the Communication subscale. For the assistant teacher, this interaction represented the largest proportion of explained variance (30.2%) with an additional 6% explained by the occasion facet. All other variance components contributed minimally towards total variance.

Despite these similarities, the discrepancy between the lead and assistant teacher was much larger for the Cooperation scale than for the Communication scale. Given the enacted model with 6 students, 7 items and 4 occasions, both teachers could make accurate screening ( $E\rho^2 = .91$ , SEM = .13 for the lead teacher;  $E\rho^2 = .88$ , SEM = .14 for the assistant teacher) and progress monitoring decisions ( $\Phi = .91$ , SEM = .13 for the lead teacher;  $\Phi = .88$ , SEM = .14 for the assistant teacher).

**Reduced model D studies.** In the final series of D studies, the item and occasion facets were jointly manipulated to determine the optimal assessment conditions to obtain adequate reliability-like coefficients for the two subscales. Results of these studies are depicted in Figures 6 and 7.

For the Communication scale, both teachers achieved adequate generalizability and dependability in as few as two to four rating occasions. The lead teacher could make reliable screening decisions using 6 items on 2 occasions, 4 items on 3 occasions, or 2 items on 6 occasions. Slightly more occasions were necessary for reliable progress monitoring decisions: 6

occasions. If three or less items were used, the number of occasions required disproportionately increased. For example, with 3 items, 13 occasions were required for adequate dependability. The assistant teacher could make reliable screening decisions with 4 items on 3 occasions, 3 items on 4 occasions, or 2 items on 6 occasions. For progress monitoring decisions 4 items on 3 occasions, 3 items on 4 occasions, or 2 items on 8 occasions were required.

For the Cooperation scale, both teachers achieved adequate dependability and generalizability with feasible assessment procedures. The lead teacher could make reliable screening decisions using 4 items on 1 occasion, 2 items on 2 occasions, or 1 item on 3 occasions. Reliable progress monitoring decisions could be made using at least 5 items on 1 occasion, 3 items on 2 occasions, or 2 items on 11 occasions. In contrast, the assistant teacher required more items and occasions to reach the .80 criterion. Accurate screening decisions required 6 items on 5 occasions, 4 items on 6 occasions, 3 items on 8 occasions, or 2 items on 11 occasions to achieve adequate generalizability coefficients. Accurate progress monitoring decisions required all 6 items on 6 occasions, 5 items on 7 occasions, 4 items on 9 occasions, or 3 items on 13 occasions.

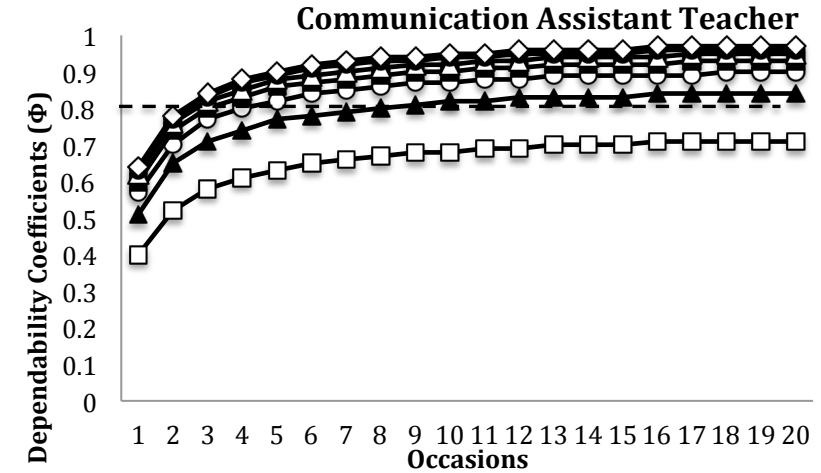
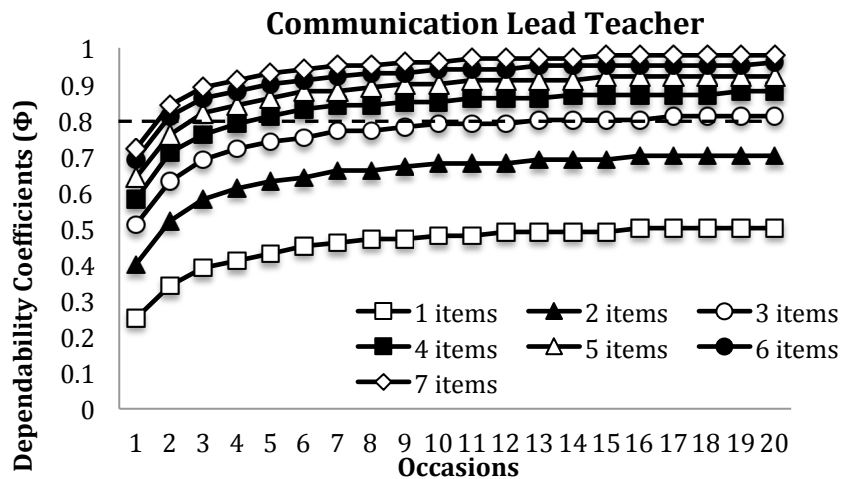
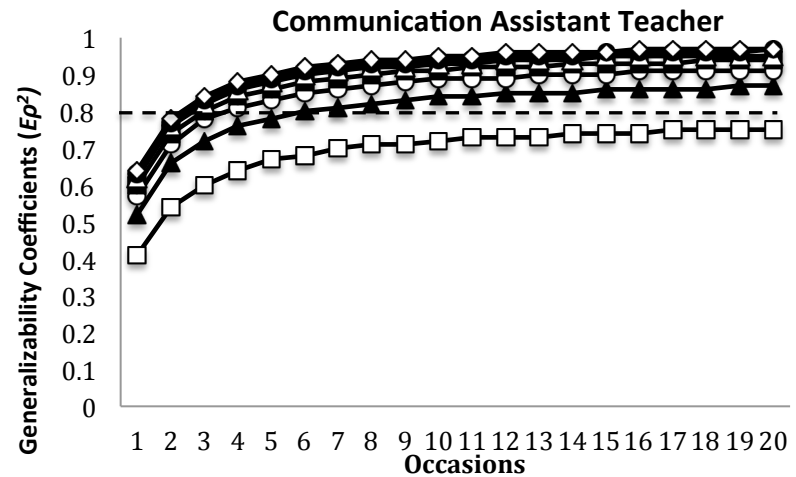
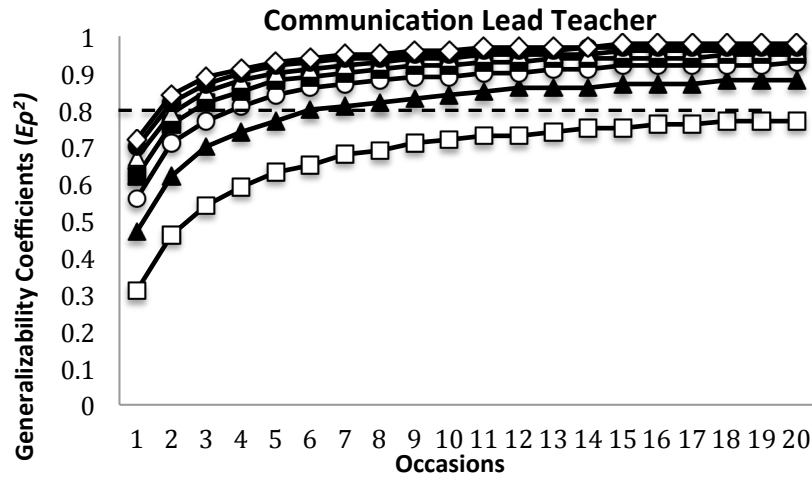


Figure 6. Generalizability and Dependability Coefficients for the Communication Subscale. The dashed lines indicate the  $\geq .80$  criterion for adequate generalizability and dependability.

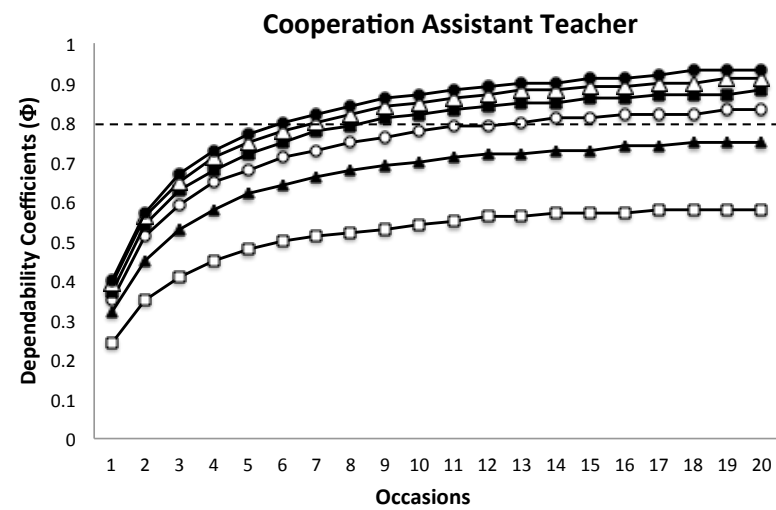
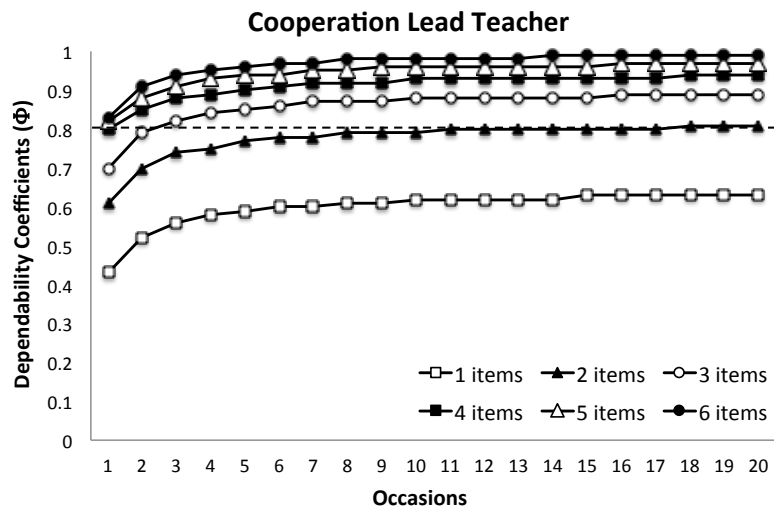
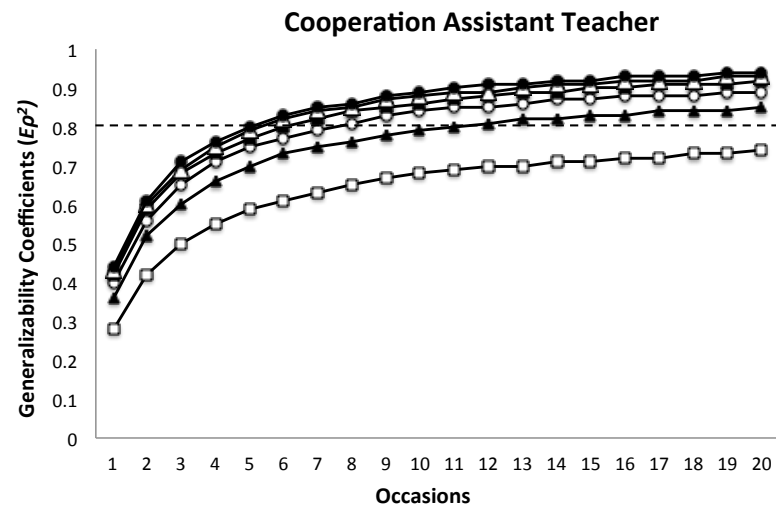
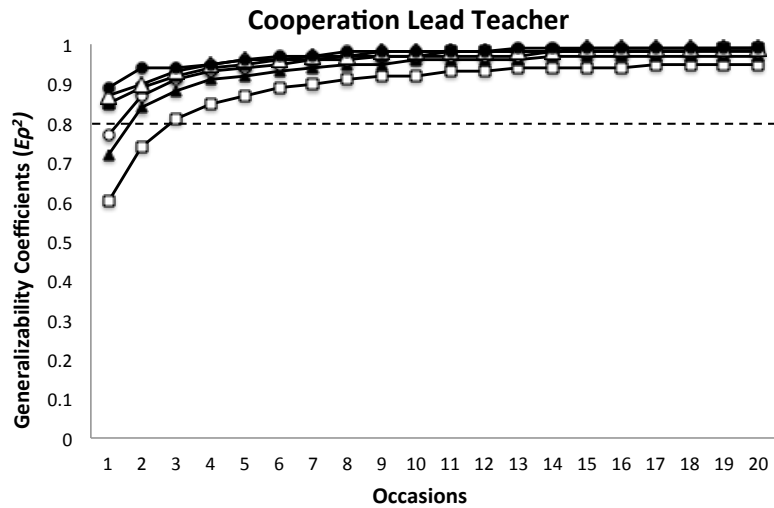


Figure 7. Generalizability and Dependability Coefficients for the Cooperation Subscale. The dashed lines indicate the  $\geq .80$  criterion for adequate generalizability and dependability.

## DISCUSSION

The original goal of this project was to develop and evaluate a set of BBRS that corresponded to important social skills domains, as the first steps of developing a tool to progress monitor social behavior. However, when the full-item subscales were explored first, results indicated that reliable decisions could not be made even with 100 rating occasions and all of the extant items from the SSIS-RS subscales. These results could largely be attributed to rater-related effects, which represented a third (Cooperation) to a half (Communication) of the total variance in the obtained scores. Therefore, reducing the number of items would negatively impact generalizability and dependability. Instead, these findings suggest that future efforts should be directed toward improving the assessment procedure to reduce unwanted variability.

The following section provides an interpretation of the results and the implications. In particular, the rater-related variance finding will be reviewed in the context of previous literature. Suggestions will be made to control for unwanted variability between raters. Based on the results and literature review, suggestions will be made to guide future efforts to develop tools to progress monitor social behavior. Finally, limitations will be discussed and suggestions will be made to improve future research.

### Summary of Statistical Findings

**Full model G and D studies.** When developing G Theory, Cronbach (et al., 1972) envisioned that G studies would be integral to the development of new measures and assessment procedures. G theory offers exceptional utility when developing new measures because the influence of several sources of variance (and their interactions) can be evaluated at once. This process can reveal undesirable sources of variability that detract from the measurement of true

individual differences. Importantly, this information can be applied to improve the design and use of assessment procedures (Cardinet et al., 2009; Cronbach et al., 1972).

Once identified, two methods can be used to control unwanted sources of variance (Cronbach et al., 1972). The first method entails modifying the measure itself (e.g. improving items, operational definitions, etc.) or further specifying operational procedures (e.g., clarifying coding procedures, rater qualifications, etc.). In this method, the revisions are made subsequent to identifying the disrupting source of variance in the initial G study, and then new data are collected and evaluated in G and D studies. The second method entails adding additional levels of the disrupting facet, which is then evaluated in D studies using the original data. Although the second method is more typically employed by the studies within in this literature review, Cronbach described the first method as preferable:

Generalizability studies ought to be regarded as part of instrument development, and therefore G studies should take place prior to the collection of the D data ... To be sure, one will occasionally use the actual D data for an analysis of generalizability. But since it is by then too late to take advantage of the information to improve the D data, this is a weak use of the method. (Cronbach et al., 1972, p. 18).

As will be discussed below, we found the first method to be necessary.

**Full model G studies.** Sources of variability for the Communication and Cooperation subscales were explored in the full model G studies. As explained previously, the full item scales were used in order to maximize obtained information. Overall, the proportion of total variance attributed to rater-related effects was larger than anticipated, whereas the proportion of variance attributed to person was smaller than anticipated. As expected, variance due to occasions was negligible, and there was some variability associated with the items facet for the Cooperation subscale but not for the Communication subscale.

The major finding from the G studies was that rater-related effects accounted for about a third (38.8% Cooperation) to nearly a half of the total variance (42.2% Communication) in the teachers' ratings. The main effect for rater indicated a small difference in rating leniency between the teachers (5.5% Communication, 9.2% Cooperation). Visual inspection of the ratings revealed that the lead teacher assigned higher ratings (i.e., skills were performed more frequently) than the assistant teacher. There is precedence for this finding; in the interrater reliability study in the SSIS-RS manual, Gresham and Elliott (2008) reported that primary teachers rated students higher on the social skills scale and lower on the problem behavior scales relative to secondary teachers. Gresham and Elliot suggested that possible qualitative or quantitative differences in teacher-student interactions may produce improvements in student behavior in the presence of the lead teacher. A similar scenario may have been in effect in the current study; the assistant teacher may have had less rapport or instructional control with the students, which may have produced different patterns of student-teacher interactions across the two teachers. In line with this explanation, the assistant teacher had recently returned from an extended leave of absence, which may have influenced her interactions with the students.

Most of the rater-related effects were attributed to interactions between the rater facet and other facets, again highlighting an important advantage of G theory relative to other measurement theories. First, the person by rater interaction (12.3% Communication, 8.4% Cooperation) indicates some evidence of a dyad-specific rater bias, which occurs when some raters are more severe when rating particular students (Hoyt, 2000). Similar to Briesch et al. (2010), visual comparisons of the teachers' ratings (see Figure 3) revealed that the teachers rated most students in a similar manner, but were discrepant for a few students. Taken together, the high percentage of variance associated with the rater and person by rater interaction components

suggests that teachers may make different decisions about individual students (Chafouleas et al, 2010). The rater by item interaction was also notable, especially for the Communication subscale (15.7%; 4.7% Cooperation), as this may reflect differences in how teachers interpreted the items. Finally, the person by rater by occasion interaction (4.7% Communication; 12.6% Cooperation) was interesting as it revealed differences in the teachers' ranking of students across occasions.

In contrast, universe score variance (i.e. person and the person by item interaction, as items was modeled as a fixed facet) was lower than anticipated (15.7% Communication, 26.8% Cooperation), given the careful specification of the universe of admissible observations. However, universe score variance has differed widely across previous studies, which may result from differences in target behaviors, populations, or study design (Volpe & Briesch, 2012).

Similarly, the person by occasion interaction (7.2% Communication, 2.8% Cooperation) was lower than expected given results reported in other studies, although the larger person by rater by occasion interactions (4.7% Communication, 12.6% Cooperation) may account for some of this variance. This indicates that these behaviors were fairly consistent across occasions. These target behaviors may be more stable than other behaviors. Alternatively, teachers may not be as sensitive to detecting variations in these behaviors, which would not be desirable for progress monitoring.

**Full model D studies.** Given the enacted model, the generalizability and dependability coefficients for the Communication ( $E\rho^2 = .62$ ;  $\Phi = .55$ ) and Cooperation subscales ( $E\rho^2 = .79$ ;  $\Phi = .68$ ) were well below the recommended  $\geq .80$  criteria for screening and progress monitoring decisions (Saliva et al., 2010). Moreover, when D studies were conducted for one



rater, coefficients decreased precipitously ( $E\rho^2 = .47$ ;  $\Phi = .40$  Communication;  $E\rho^2 = .46$ ;  $\Phi = .39$  Cooperation). These results indicate that the teachers' ratings are not interchangeable.

After identifying rater-related effects as a disruptive source of variance, both methods to control unwanted variance (Cronbach et al., 1972) were considered. The second method was evaluated using the extant data to examine the impact of additional raters in a series of D studies. These results were compared with independent manipulations of the occasions and items facets (see Figure 4). Finally, the number of raters, items and occasions were systematically manipulated to determine the minimum number of raters required to obtain adequate generalizability and dependability coefficients. At minimum, adequate generalizability required 5 raters with 6 items and 9 occasions for Communication, and 2 raters, 3-5 items and 5-8 occasions for Cooperation. Adequate dependability required more raters: at least 7 raters, 6 items and 10 occasions for Communication and at least 4 raters, 5-6 items, and 5-7 occasions for Cooperation. Furthermore, results indicated that reducing the number of items on the scale would negatively impact generalizability and dependability coefficients unless offset by adding additional raters.

While results indicated reliable decisions could be made by adding additional raters, the number of raters required to do so would be untenable in typical school settings. Furthermore, as we were primarily interested in creating a measure suitable for progress monitoring purposes, the dependability coefficients would be used to make absolute decisions. As described above, this would require more raters than would be necessary for screening decisions.

Thus, we were not able to reasonably control for rater-related effects using a feasible assessment procedure. Alternatively, changing the items and assessment procedures, and then collecting new data to evaluate these revisions with a new series G and D studies might control

this variance. These procedures were beyond the scope of the current study. However, the results the G and D studies have important implications for future research and scale development. Therefore, our discussion will include recommendations to improve the scale and the assessment procedures.

Before continuing, these findings must be qualified with the recognition that, as with all G theory studies, results are highly dependent on the definition of the universe of generalization. In the current study, the rater universe was broadly defined as all preschool teachers with similar training and experience, to develop a scale that could be widely used in preschool classrooms. If another user was only interested in the classroom in this study, the rater facet could be modeled as fixed facet. By defining this facet (which represented a substantial proportion of the total variance) narrowly, sizable improvements in generalizability and dependability coefficients ( $E\rho^2$ ,  $\Phi = .89$  Communication;  $E\rho^2 = .93$ ,  $\Phi = .92$  Cooperation) would be achieved. These results would then support development of abbreviated scales. Christ et al. (2010) reported similar results when raters were modeled in as a fixed facet in D studies based on G study components with sizable rater-related variance effects.

Likewise, different results would be obtained if the definition of the item universe was changed. In the current study, item was modeled as a fixed facet that consisted of the extant items on the Communication and Cooperation subscales from the SSIS-RS. However, another researcher might interpret these items as samples from a larger population of similar items, and may thus choose to model items as random facet. We did not evaluate this scenario, as the rater by item interaction suggested differences in the teachers' interpretation of the items. As will be discussed shortly, controlling for this source of variance may require that contextual information

be added to items. These alterations would require a change to the universe of admissible observations such that the extant items may no longer represent a sample of this universe.

**Reduced model G and D studies.** As the original purpose of this study was to evaluate an abbreviated rating scale, it was determined that examining sources of variability obtained by individual teachers may yield useful information for future scale development. Therefore, a series of reduced model G studies were conducted to analyze the ratings of each teacher separately (see Tables 8 and 9). Interpreting variance within rater generally produced large gains in the universe score variance. This may suggest that individual raters are better at detecting behavioral differences between students. However, given the percentage of person by rater variance in the full model, some of this improvement may be explained by rater biases for some students (i.e., dyad-specific rater bias). Interestingly, across both scales, the assistant teacher rated student behavior as being more variable across occasions than the lead teacher. This difference may likely impact progress monitoring decisions and further speaks to the need to understand factors that may contribute to differences between raters.

Finally, it is notable that across both scales, the lead teacher's ratings indicated that (across students) some items were performed more frequently than others. In contrast, the assistant teacher rated overall behavior as more variable (e.g. changes in the rank-ordering of students across days), yet rated less differences between items. Overall, these differences indicate that the teacher's ratings were likely influenced by different factors. Again, this highlights the importance of developing procedures to control for unwanted variability between raters.

Based on the improved universe score variance in the reduced model G studies, it was anticipated that reliable decisions might be achieved when the individual teachers completed the

ratings. Thus, reduced model D studies were conducted to determine the number rating occasions and items required for each teacher to achieve this. Across both scales and teachers, this could be obtained with range of feasible assessment options.

Although the reduced model G studies indicated that the teachers' ratings might have been influenced by different factors, it is encouraging that reliable ratings were obtained for both teachers using feasible assessment procedures. Moreover, results supported a reduction the number of items and ratings occasions. However, it is important to recognize that these results should only be considered to provide preliminary support for using multiple-item scales over single item scales. Conducting G and D studies within rater restricts the universe of generalization to the individual teachers in this study, such that these results may not generalize to other teachers. Therefore, the scales and/or assessment procedure should be first revised in order to control for substantial rater-related effects.

### **Implications of Findings**

Although we were unable develop an abbreviated rating scale suitable for progress monitoring social behavior with the extant items and instructions from the SSIS-RS, it should be recognized that scale development is typically a process, replete with multiple iterations and revisions. Thus, the current study extends the research in several important ways. First, there are only a handful of G theory studies that have investigated methods to assess social behavior in schools; this is the second study to attempt evaluation of a BBRs. Next, the study design allowed for more in depth analysis of rater related effects than in previous studies. The current study also added preliminary support to Volpe et al.'s (2011; Volpe & Briesch, 2012) line of work that demonstrated improved efficiency of multiple item scales over single-item scales.

Finally, differences were noted between the Communication and Cooperation subscales, as well as other target behavior from previous studies.

**Rater effects.** As previous research on the extent of rater related-effects in the assessment of social behaviors has been equivocal, an in depth analysis of these effects was warranted, particularly as this facet was not analyzed in previous BBRs studies. In this current study, rater-related effects accounted for 42.2% of total variance for the Communication scale and 38.3% for Cooperation. Of particular concern, the large proportion of variance associated with the rater, person by rater, and rater by item components suggested that teachers may make different decisions about individual students (Chafouleas et al, 2010). If a rater is biased when rating particular students, this may result in adverse outcomes for these students. This should be considered as a potential negative consequence of test use (Messick, 1995) that may arise when information is provided by a single rater.

In comparing results across studies, an interesting pattern emerged. Rater-related effects were more pronounced when ratings were conducted by teachers (about one-fourth to half of the total variance in the previous literature) than by graduate student observers (0-12% of total variance), which have not gone unnoticed by other researchers (Chafouleas et al., 2010; Volpe & Briesch, 2012). While this would seem to indicate that trained observers are more accurate reporters, SDO is not without its limitations (Gresham, 2011). Moreover, reliance on school psychologists limits the feasibility of assessment procedures, particularly within an RTI-context.

Several factors may contribute to this effect. First, teachers in these studies observed students while they taught, whereas graduate students focused solely on observations (Briesch et al., 2010; Chafouleas et al., 2010; Volpe & Briesch, 2012). Next, teachers simultaneously observed more students than graduate students. Together, these factors may result in teachers

missing more instances of behavior, particularly when observing low base rate behaviors that are not disruptive to the classroom ecology (Hersh & Walker, 1983). Finally, graduate students (and school psychologists) typically have advanced training and experience in conducting behavioral assessments (Riley-Tillman et al., 2008), which may make them less susceptible to rating biases.

**Scale length.** As previously discussed, a major advantage of BBRS is that these scales have been reported to retain the acceptable psychometric features from the original scales with fewer items (Gresham et al., 2010; Volpe & Gadow, 2010; Volpe et al., 2011), which is advantageous for progress monitoring purposes. Furthermore, the BBRS study by Volpe et al. and the multiple-item DBR study by Volpe & Briesch (2012) suggested that brief, multiple-items scales were more efficient (i.e. required less rating occasions) than similar single item scales. Thus, BBRS appear to be a promising measurement methodology.

Unlike like the previous BBRS studies, we were unable to develop an abbreviated scale with adequate levels of reliability while retaining the items and rating procedures from the original scale. This likely reflects differences in study design, as the previous studies did not investigate raters as a source of variance. For example, in the Volpe et al. (2011) study, each student was rated by a separate informant. In G theory terminology, this describes a hidden facet, which means that the relatively high percentage of variance attribute to persons may have been confounded by rater-specific effects (Cronbach et al., 1972; Webb et al., 2006).

In the current study, when the ratings by individual teachers were examined in the reduced model D studies, results were consistent with previous studies (Volpe et al., 2011; Volpe & Briesch, 2012) in that fewer rating occasions were required to achieve adequate generalizability and dependability when multiple items were used. This may appear intuitive given the Spearman Brown prophecy. However, in this study and the Volpe et al. study, the

items facet produced a point of diminishing returns beyond which additional items produced little effect on generalizability and dependability coefficients. Together, the results from this study, Volpe et al. 2011, and Vople and Briesch (2012) supported the efficiency of multiple items scales. This is notable since each study investigated different target behaviors, populations and assessment methods (BBRS and multiple-item DBRs). Thus, this appears to be a robust finding. Overall, brief, multiple-items appears to be a promising methodology that should continued to be explored in future studies.

**Target behaviors.** Generalizability and dependability coefficients may also be dependent on the target behavior. In the current study, the universe score variance, and generalizability and dependability coefficients were higher for the Cooperation subscale than for Communication. This finding may indicate that there is more variability between students for some behaviors than others. Alternatively, teachers might be more attuned to individual differences in cooperation than other domains. Given the contextual nature of social skills, cooperation may be more readily observed by teachers as these behaviors are relevant to student-teacher interactions, whereas communication behaviors may be more relevant to peer interactions (i.e., teacher and peer-preferred social behavior, see Walker, Irvin, Noell, & Singer, 1992). Moreover, teachers consider cooperation to be among the most important social skills, which may make deficits in this area more salient to teachers (Gresham & Elliott, 2008).

Other studies have also reported differences across various behaviors. Several studies have reported that raters overestimated the frequency of disruptive behaviors, and underestimated the frequency of desirable behaviors (Chafouleas et al., 2012; Christ et al., 2010). However, differences across studies have been reported even when using same target behaviors.

This likely indicates that other factors, such as populations and level of specification may impact results (Chafouleas et al., 2007; 2010; Riley-Tillman et al., 2009).

When reviewing these findings, it is important to consider that DBRS and BBRS are assessment methods, rather than a single measure. As with CTT, it is important to establish the generalizability and dependability for measures separately across target behaviors, populations, and levels of behavioral specificity (Chafouleas et al.; Volpe & Briesch, 2012). Next, we will discuss our findings in light of previous research.

### **Relevance of Current Research to Previous Literature**

Perhaps the most notable finding in the current study was that rater-related variance affected generalizability and dependability coefficients to the extent that an abbreviated scale was not justified. Since the rater-related effects were much higher than anticipated, a review of previous literature on rater reliability is warranted.

The informant discrepancy literature bears relevance in interpreting the results from the current study. This literature contains one of the most robust findings in clinical psychology: correlations between informants yield low to moderate levels of agreement (De Los Reyes & Kazdin, 2005). From a CTT perspective, low correlation between pairs of informants is seen as rating error, low reliability, or indication that one rater is a more accurate reporter than others (Wright, Zakriski, Hartley & Parad, 2011). However, behavioral researchers contend that these so-called disagreements reflect important contextual differences in behavior, particularly in regard to the situational specificity of behavior (Achenbach et al, 1987; De Los Reyes, 2011).

Supporting this assertion, it is well established that informants in similar settings yield higher correlations than informants in different settings (Achenbach et al, 1987; Gresham et al., 2010). Particularly relevant to the current study, on a recent cross informant study of the SSIS-



RS, correlations between pairs of teachers were  $r = .63$  for the Communication subscale and  $.60$  for Cooperation subscale whereas corresponding teacher-parent correlations were  $.28$  for both subscales. This corresponds to 40% shared variance for the Communication subscale and 38% for Cooperation subscale, remarkably similar to the percentage of variance attributed to rater-related effects in the current study.

De Los Reyes and Kazdin (2005) describe three factors that emphasize the importance of studying informant discrepancies. First, there is no universal standard to indicate the presence or absence of a disorder. Similarly, there is no standard for determining which informant provides the most accurate information. Informants have different motivations and tolerances that may influence judgments (Chafouleas, Kilgus, Riley-Tillman, Jaffrey, & Harrison, 2012; Gresham, 2011; Shinn, Tindal & Spria, 1987). Thus, in the current study, it would be inappropriate to state that the lead teacher rated the students more accurately than the assistant teacher. Second, attempts to force concordance between informants by confronting them about discrepancies have generally not led to improved agreement (Nguyen et al., 1994), but may pressure informants to produce similar, rather than accurate ratings (Angold et al., 1987). Last but not least, discrepancies often are indicators of situational or contextual factors that highlight relevant clinical information about the child's functioning in different settings (Achenbach et al., 1987). For example, discrepancies in the teachers rating in this current study could reflect student behavioral differences that arise due to differences in the teachers' classroom management styles. Such information could prove invaluable to intervention planning efforts.

While the cross-informant typically discusses raters in different settings, it also pertains to raters in the same setting, such as the teachers in the current study. As suggested by Achenbach et al. (1987) although informants in the same setting are exposed to similar samples

of behavior, these samples are rarely identical (also see Dirks, De Los Reyes, Briggs-Gowan, Cella, & Wakschlag, 2012). In the current study, the two teachers likely had some different interactions with the students, such as during small group activities. This may in part explain differences in variance partitioning for the two teachers in the reduced model G studies.

A variety of factors have been identified as potential mediators of informant discrepancies. These include child characteristics such as age, gender, race, characteristics of the presenting problem (externalizing and observable behaviors associated with higher agreement), and informant characteristics such as training and previous experience (Achenbach et al., 1987; De Los Reyes & Kazdin, 2005; Hoyt & Kern, 1999). Many of these variables have been suggested as contributing to differences in generalizability and dependability coefficients across the previously G theory (Chafouleas et al., 2010; Vople & Briesch, 2012)

De Los Reyes and Kazdin (2005) provide a theoretical framework for explaining informant discrepancies based on research from the social-cognitive literature, which they have termed the ABC model. To summarize, the ABC model states that in addition to contextual factors, informant discrepancies arise when the informants make different attributions regarding the cause of children's behavior (Jones & Nisbett, 1971) and have different perspectives regarding the goals of assessment or the importance of treating individual target behaviors.

Although the ABC model is traditionally applied to pairs of informants in different settings, the ABC theory is also relevant to informants in the same setting (De Los Reyes & Kazdin, 2005). For example, the two teachers in the current study may have different perspectives on what behaviors most warrant intervention for different students, which may have produced different memories of behavior that are consistent with their unique perspectives (Gresham & Elliott, 2008).

The ABC model may also provide a conceptual framework for investigating methods to reduce rater bias. An interesting avenue for future research would be to systematically manipulate factors that may contribute teacher's reliance on heuristics when rating social behavior. De Los Reyes and Kazdin (2005) suggested that this might be achieved by the inclusion of contextual questions (especially regarding the presence or absence of behavior under specific conditions) into assessment procedures. This may encourage the use of systematic memory retrieval processes and decrease reliance on heuristic memory processes associated with rating biases (Tversky & Marsh, 2000). Applied to the current study, two areas that merit further exploration includes the potential effects of cognitive load on teachers rating accuracy (e.g. the number of students observed, other activities performed by teachers during observations), and the length of rating intervals.

### **Recommendations to Address Rater Biases**

In order to develop more reliable assessment procedures, the current study suggests that rater biases must be reduced. Thus, this section will review methods that may be used to mitigate rater-related errors. While the reduction of rater bias is desirable, it is not desirable to completely eliminate differences between raters as these differences may reflect important contextual information, such as the situational specificity of behaviors (Achenbach et al., 1987; De Los Reyes, 2011), patterns of interactions unique to rater/individual dyads (Hoyt & Kern, 1999), or differences in the rater's judgments of the social importance of behaviors (Gresham & Elliott, 2008; Gresham, Sugai, & Horner, 2001). It is these differences that make data collection from multiple sources an important part of evidence-based assessment practice (Hunley & Mash, 2007; Merrel, 2003).

**Rater training.** One way to improve consistency between raters would be to develop rater training procedures. These procedures are used to reduce rater biases/error by ensuring that raters are able to: (1) recognize target behaviors, (2) use the rating system with ease and (3) conduct ratings in reference to some pre-established standard (Spool, 1978). Most of the rater training literature within the field of school psychology has focused on training school psychologists in observational procedures (Lebel, Kilgus, Briesch, & Chafouleas, 2010). In practice, teachers typically complete BRS after reading brief written instructions as BRS manuals do not usually include rater training procedures (Floyd & Bose, 2003). In keeping with this practice, in the current study, teachers were provided with written instructions, a brief review with the research assistant and an opportunity to ask questions about the rating procedures. While minimal training requirements are generally cited as strength of BRS, lack of training may unduly contribute to rater error (Elliott et al., 1993; Volpe et al., 2009).

Although there has been minimal research in this area for BRS or BBRS, a recent series of studies has investigated rater-training procedures with DBRS (Chafouleas, et al., 2012; Lebel, et al., 2010; Schlientz, Riley-Tillman, Briesch, Walcott, & Chafouleas, 2009). This research suggested that the optimal level and type of training may be dependent on the target behavior or rate of behavior. Overall, indirect training procedures (e.g. vignettes, instructional videos) have been demonstrated to be less effective than direct skills training packages (e.g. modeling, rehearsal and performance feedback; Chafouleas et al., Chafouleas, et al., 2005; Lebel et al.). Additionally, for disruptive behaviors that occur at high rates, additional training methods, such as frame of reference or rater error training may be warranted (see Chafouleas et al., 2012 for a review).

DBR researchers have found that 30-60 minutes of training produced improvements in rater accuracy (Harrison, Jaffrey, Johnson, Riley-Tillman, Chafouleas, & Christ, 2011). Interestingly, Chafouleas et al. (2012) found that moderate training (three practice opportunities with modeling and feedback) was more effective than more extensive training (e.g. six practice opportunities, additional of frame of reference and rater error training) when rating disruptive behavior that occurs at low rates. Chafouleas (et al., 2012) suggested that at times, additional training might occasionally confuse raters.

On the whole, these results suggested that the rating accuracy of inexperienced raters is improved after 30-60 minutes of direct training (Chafouleas et al., 2012; Harrison et al., 2011). This is an important finding when weighing the feasibility of different assessment procedures. Although expert raters appear to be less susceptible to rater error (Chafouleas et al., 2010; Volpe et al., 2012) for frequent progress monitoring purposes, it may be more feasible to have teachers complete behavior ratings after attending a brief training. This appears to be fruitful area for future research when developing BBRS to assess social behavior.

**Operationalization of items and procedures.** As discussed previously, the rater by item interaction in the current study was larger than anticipated (15.7% Communication, 4.7% Cooperation). This suggests that the teachers interpreted the items differently, particularly on the Communication scale. Improving the operationalization of items and rating procedures may help reduce these effects. First, scales may benefit from the inclusion of more observable items. Supporting this assertion, meta-analyses of cross-informant rating have generally yielded higher interrater correlations for externalizing behaviors than for internalizing behaviors (Achenbach, Krukowski, Dumenci, & Ivanova, 2005; Achenbach et al., 1987). Furthermore, a meta-analysis of rater effects reported in G theory studies found that measures with high inference items were

associated with high levels of rater-related variability (nearly 50% of total variance), whereas explicit measures, such as frequency counts, the variance attributed to rater effects were minimal (Hoyt & Kerns, 1999).

Reduction of variance related to differences in item interpretation may be also achieved by providing scoring guidelines to operationalize ratings (Volpe & Hintz, 2009; also the Direct Observation Form, McConaughy & Achenbach, 2009 is an excellent example). Cronbach et al. (1972) recommended that in addition to these procedures, an important part of the specification of the universe of admissible observations is the operationalization of assessment procedures. This might include description of instruction provided to raters, rating rules (e.g. rubric or behavioral anchors), qualifications of raters, as well as any rater training procedures (Cronbach et al).

**Multiple raters.** Given similar findings regarding the influence of rater-related effects, previous researchers have advocated that DBR ratings should be conducted within-rater only until suitable rater training procedures have been developed (Chafouleas et al., 2010; 2012). However, the high proportion of rater variance found in this and other studies (Chafouleas et al., 2007; Christ et al., 2010) warrants concern when decisions are made with information from one informant only, particularly if the data is used to support high-stakes decisions. Although the results of the current study suggest that a single rater can produce excellent reliability-like coefficients using relatively few items and occasions, both systematic (rater specific) and rater by person (dyad-specific) rater errors may result from raters making different decisions about students and is contrary to best practices in assessment (Hunley & Mash, 2007). Certainly, these researchers have recognized this limitation and continue to explore rater-training procedures to

control for rater biases (Chafoulaes et al., 2012; Lebel et al., 2010; Harrison et al., 2011; Schlientz et al., 2009).

Clearly, there are methodological as well as clinical advantages to using multiple raters (De Los Reyes & Kazdin, 2005; Dirks et al., 2012; Hoyt, 2000). In the current study, dependability and generalizability coefficients improved substantially as additional raters were added to the assessment procedure. Yet, while optimal methods to assess cross-informant discrepancies have been identified, there is no consensus on how to meaningfully aggregate information across raters. As aptly stated by Gresham (2011):

... (a) Multiple sources of information are often used to assess students' social behavior without guidance as to which source of information to trust or weight most heavily; (b) use of a single source of information will necessarily restrict the conclusions and recommendations to be drawn; and (c) the use of single or multiple sources of information in research studies often significantly changes the conclusions that might be drawn from about an individual (Gresham, 2011, p. 276).

A major limitation of aggregation strategies is that important contextual information from different raters may be obscured (De Los Reyes et al., 2011; Dirks et al., 2012). This issue is quite complex as there are many factors that influence informants' ratings and each assessment situation is likely to have a unique constellation of these factors (Achenbach, 2011; Kraemer et al., 2003). In summary, while research indicates that multiple raters should be used to assess social behavior, it is not yet clear how to meaningfully combine this information across raters to make progress monitoring and other important decisions.

### **Recommendations for Future Scale Development**

Although abbreviated rating scales were not developed as intended, results from this study and its literature review offer insight for future development of BBRS for social behavior. First, new items might be developed to include more observable behaviors or reflect contextual information. It may also be helpful to assess the intensity, as well as frequency of behaviors,

particularly when rating problem behaviors (Gresham, 2011). An expert panel of social skills researchers and a thorough literature review should prove useful in developing a pool of potential items. Traditional scale construction methods such as factor analysis may also be helpful in reducing the item pool.

Once this pool is developed, another G study should be conducted to analyze sources of error variance. The EduG software offers a tool called G Facet analysis in which items (or specific levels of any facet) may be iteratively deleted to see how each item influences generalizability and dependability. This analysis may also reveal factors that contribute to rater-related effects. After further refining the item pool, a preliminary series of G and D studies may be conducted using all the items in the refined pool, as this may help determine the number of items and occasions for inclusion on an abbreviated scale. Using this information, an abbreviated scale may be developed using one of the methods described previously.

While we did not complete construction of the abbreviated scales as intended using factor analytic construction methods and the recommendations of an expert, we encountered a problem with the factor analytic approach when initially exploring this method. The Communication scale loaded into two factors and it was unclear how to select the items with the highest factor loading. An alternative method would be to develop a standard set of items using the change-sensitive approach (Gresham et al., 2010; Meier et al., 2008; Volpe et al., 2010). As noted by Gresham et al., this approach is quite promising as could also provide evidence of change-sensitivity, the second stage of developing progress monitoring tools (Fuchs, 2004). As recommended by Volpe and Briesch (2012), a range of tools should be developed to assess both specific and global objectives. Thus, it may be beneficial (as originally intended with this study)



to develop a full set of BBRS to correspond with specific social skills domains so that an individualized approach to assessment could be employed while using reliable measurement.

Another intriguing avenue to consider in developing rating scales is how rating interval durations affect generalizability and dependability. So far, results have been mixed. For example, David-Ferguson, Briesch, Volpe, & Daniels (2012) found that two 30-minute or five 10-minute observations of for academic engagement produced similar results. In contrast, an earlier study by McWilliam and Ware (1994) found that fewer observations with longer durations were more efficient than more observations with shorter intervals.

Most G theory studies of DBR and SDO have employed brief intervals (e.g. 30 minutes or less) and measured globally defined behaviors. However, behaviors assessed on BRS and BBRS may be of interest due to their intensity rather than their frequency. These behaviors may not be observed on a daily basis. Yet, the G theory studies using these methods have all used daily interval durations. As halo rating effects are more likely to occur in the absence of adequate samples of the target behavior (Feeley, 2002), future studies may benefit from investigating different rating interval lengths.

### **Limitations**

In addition to the areas discussed thus far, some additional limitation merit further discussion. First, although 14 components were specified in this study, some variance was still unexplained by the model (31% Communication, 16.8% Cooperation). While this is comparable, or an improvement over similar studies, additional facets may need to be specified to more fully account for the total variance.

Next, while 624 total data points were collected (e.g., 6 students x 2 raters x 6-7 items x 4 occasions for two scales) the number of students and conditions sampled in the enacted model

were small, albeit similar to sizes used in previous studies. Although most facets were modeled as random, there may be contextual features present in this classroom that may not have been adequately addressed in our specification of the universe of admissible observations. For example, some items may be opportunity dependent, and thus, behaviors may occur more in some classrooms than others regardless of specification of items, rater or assessment procedures (R. J. Volpe, personal communication, April 17, 2012).

As noted by Smith (1981), the stability of G theory variance component estimates improves as more data points are added. However, obtaining stability with a different design was carefully weighed against the advantages of a fully crossed design to the interpretability of variance components (especially rater effects) as well as a desire to reduce teachers' rating load. This trade off has been referred to as "the Achilles heel of G theory" (Shavelson & Webb, 1981, p. 138). One interesting strategy proposed by Smith (1981) would be to estimate G study components with a carefully constructed series of nested of G studies to serve as "mini replications" of facets (p. 153). Using this technique, data could be collected from multiple classrooms and schools without losing interpretability of facets.

Next, the study population of preschool students at risk for social skill deficits necessarily restricts generalization of results to other populations. However, this definition appeared to be warranted given that progress monitoring tools for social behavior would be primarily used for students with social-behavioral deficits. As expected, average obtained scores in this study were lower than in the SSIS-RS standardization sample (Gresham & Elliott, 2008). Given lower of rates of social behavior, a floor effect may be in place such that the 4-point scale used in this study could not adequately capture behavioral differences between individuals. This may

explain why lower percentages of variance were explained by *person* in this study than in previous studies without this population restriction.

Thus, caution is warranted in generalizing these findings to a wider population of preschool students. This may hamper potential use of these scales as screening tools. This is unfortunate, given that this may be an ideal use of these scales given that the obtained generalizability coefficients were larger than dependability coefficients. However, as new coefficients would be calculated after changes were made to improve the scale or assessment procedures, this conclusion may be premature and should be reevaluated in future research.

As discussed above, many of the limitations relate to the specification of the universe of admissible observations and universe of generalization. However, this should be characterized as a strength of G theory as it requires careful attention to the design and use of assessment procedures (Brennan, 2010). Furthermore, as it appears that results are contingent on target behaviors, population and measurement methods, it is critical to continue systematic research to understand the influence of these variables on the assessment of social behavior.

## CONCLUSION

Although the intent of this project was to develop and investigate abbreviated rating scales corresponding to social skills domains on the SSIS-RS, in execution, this investigation led to the exploration of full-item subscales themselves. As this study illustrates, G theory provides an excellent methodology for school psychologists to develop better assessment measures and to evaluate the optimal conditions for assessment. While our results precluded the abbreviation of the subscales, preliminary results from the reduced model D studies do lend support to the efficiency of multiple-item scales.

In the current study, rater-related effects accounted for a large proportion of the total variance in the rating scales. This produced detrimental effects on generalizability and dependability coefficients. Using the extant items and rating procedures from the SSIS-RS, adequate generalizability and dependability coefficients could not be obtained using assessment procedures that would be feasible in typical school settings. These findings highlight the need to develop procedures to reduce rater bias, such as rater training, and operationalization of items and rating procedures when assessing social behavior. Once revised scales are developed, further research will be necessary to establish the sensitivity and utility of these tools (Fuchs, 2004).

## REFERENCES

- Achenbach, T. (2011). Commentary: Definitely more than measurement error: But how should we understand and deal with informant discrepancies? *Journal of Clinical Child and Adolescent Psychology, 40*, 80–86.
- Achenbach, T. M., Krukowski, R. A., Dumenci, L., & Ivanova, M. Y. (2005). Assessment of adult psychopathology: Meta-analyses and implications of cross-informant correlations. *Psychological Bulletin, 131*, 361–382.
- Achenbach, T., McConaughy, S., & Howell, C. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*, 213-232.
- Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms and profiles*. Burlington, VT: Research Center for Children, Youth, and Families, University of Vermont.
- AERA, APA, & NCME. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational and Psychological Research Association.
- Alberto, P., & Troutman, A. C. (2009). *Applied behavior analysis for teachers*. Merrill/Pearson.
- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders (4th ed.- text revision)*. Washington, DC: Author.
- Angold, A., Weissman, M. M., John, K., & Merikangas, K. R. (1987). Parent and child reports of depressive symptoms in children at low and high risk of depression. *Child Psychology & Psychiatry & Allied Disciplines, 28*, 901-915. doi:10.1111/j.1469-7610.1987.tb00678.x
- Baer, D. (1977). Reviewer's comment: Just because it's reliable doesn't mean you can use it. *Journal of Applied Behavior Analysis, 10*, 117–119.
- Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1*, 91-97.
- Batsche, G., Elliott, J., Graden, J., Grimes, J., Kovalski, J., Prasse, D., Reschly, D., Schrag, J., & Tilly, D. (2005). *Response to intervention: Policy considerations and implementation*. Alexandria, VA: National Association of State Directors of Special Education.

- Bergeron, R., Floyd, R. G., McCormack, A. C., & Farmer, W. L. (2008). The generalizability of externalizing behavior composites and subscale scores across time, rater, and instrument. *School Psychology Review, 37*, 91-108.
- Brennan, R. L. (2000). (Mis)Conceptions about generalizability theory. *Educational Measurement: Issues And Practice, 19*, 5-10.
- Brennan, R. L. (2003). Generalizability theory. *Journal of Educational Measurement, 40*, 105–107. doi: 10.1111/j.1745-3984.2003.tb01098.
- Brennan, R. L. (2010). *Generalizability theory*. Springer-Verlag: NY.
- Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*, 1-21.
- Briesch, A. M., & Volpe, R. J. (2007). Important considerations in the selection of progress-monitoring measures for classroom behaviors, *1*, 59-74.
- Briesch, A. M., Kilgus, S.P., Chafouleas, S.M., Riley-Tillman, T.C., & Christ, T.J. (2012). The influence of alternative scale formats on the generalizability of data obtained from direct behavior ratings single items scales. *Assessment for Effective Intervention*. Advance online publication. doi:10.1177/1534508412441966.
- Briesch, A.M., Chafouleas, S.M., & Riley-Tillman, T.C. (2010). Generalizability and dependability of behavior assessment methods to estimate academic engagement: A comparison of systematic direct observation and Direct Behavior Rating. *School Psychology Review, 39*, 408-421.
- Busse, R. T. (2005). Rating scale applications within the problem-solving model. In R. Brown-Chidsey (Ed.), *Assessment for intervention: A problem-solving approach* (pp. 200-218). New York, NY US: Guilford Press.
- Caprara, G. V., Barbaranelli, C., Pastorelli, C., Bandura, A., & Zimbardo, P. G. (2000). Prosocial foundations of children's academic achievement. *Psychological Science, 11*, 302-306.
- Chafouleas, S.M., Briesch, A.M., Riley-Tillman, T.C., Christ, T.C., Black, A.C., & Kilgus, S.P. (2010). An investigation of the generalizability and dependability of Direct Behavior Rating Single Item Scales (DBR-SIS) to measure academic engagement and disruptive behavior of middle school students. *Journal of School Psychology, 48*, 219-246. doi:10.1016/j.jsp.2010.02.001.

- Chafouleas, S. M., Christ, T. J., Riley-Tillman, T., Briesch, A. M., & Chanese, J. M. (2007). Generalizability and dependability of direct behavior ratings to assess social behavior of preschoolers. *School Psychology Review, 36*, 63-79.
- Chafouleas, S. M., Kilgus, S. P., & Hernandez, P. (2009). Using Direct Behavior Rating (DBR) to screen for school social risk: A preliminary comparison of methods in a kindergarten sample. *Assessment for Effective Intervention, 34*, 214-223.  
doi:10.1177/1534508409333547.
- Chafouleas, S. L., Riley-Tillman, T., McDougal (2002). Good, bad, or in-between: how does the daily behavior report card rate? *Psychology In The Schools, 39*, 157-169.
- Chafouleas, S. M., Riley-Tillman, T. C., & Sassu, K. A. (2006). An investigation of the reported acceptability and usage of Daily Behavior Report Cards (DBRCs) by teachers. *Journal of Positive Behavior Interventions, 8*, 174-182.
- Chafouleas, S.M., Riley-Tillman, T.C., Sassu, K.A., LaFrance, M.J., & Patwa, S.S. (2007). Daily behavior report cards (DBRCs): An investigation of consistency of on-task data across raters and method. *Journal of Positive Behavior Interventions, 9*, 30-37.
- Chafouleas, S. M., Sanetti, L. H., Kilgus, S. P., & Maggin, D. M. (2012). Evaluating sensitivity to behavioral change using direct behavior rating single-item scales. *Exceptional Children, 78*, 491-505.
- Chafouleas, S. M., Volpe, R. J., Gresham, F. M., & Cook, C. R. (2010). Models special series: Behavioral assessment within problem-solving school-based behavioral assessment within problem-solving models: Current status and future directions. *School Psychology Review, 39*, 343.
- Christ, T. J., & Hintze, J. M. (2007). Psychometric considerations of reliability when evaluating response to intervention. In S. R. Jimmerson, A. M. VanderHayDen & M. K. Burns (Eds.), *Response to intervention handbook* (p. 93-105). New York: Springer.
- Christ, T.J., Riley-Tillman, T.C., Chafouleas, S.M. (2009). Foundations for the development and use of direct behavior ratings (DBR) to assess and evaluate student behavior. *Assessment for Effective Intervention, 34*, 201-213.
- Christ, T. J., Riley-Tillman, T. C., Chafouleas, S. M., & Boice, C. H. (2010). Generalizability and dependability of Direct Behavior Ratings (DBR) across raters and observations. *Educational and Psychological Measurement, 70*, 835-843.

- Christ, T. J., Riley-Tillman, T. C., Chafouleas, S. M., & Jaffrey, R. (2011). Direct Behavior Rating: An evaluation of alternative definitions to assess classroom behaviors. *School Psychology Review, 40*, 181-199.
- Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis*. Minneapolis, MN: Pearson Assessments.
- Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy, 8*, 411-426.
- Cone, J.D. (1978). The Behavioral Assessment Grid (BAG): A conceptual framework and taxonomy. *Behavior Therapy, 9*, 882—888.
- Crews, S. D., Bender, H., Cook, C. R., Gresham, F. M., Kern, L., & Vanderwood, M. (2007). Risk and protective factors of emotional and/or behavioral disorders in children and adolescents: A “mega”-analytic synthesis. *Behavioral Disorders, 32*, 64-77.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group/Thomson Learning.
- Cronbach, L. J., Gleser, C. G., Rajaratnam, N., & Nanda, H. (1972). *The dependability of behavioral measurements*. New York: Wiley.
- DiPerna, J., & Elliott, S. N. (2002). Promoting academic enablers to improve student achievement: An introduction to the mini-series. *School Psychology Review, 31*, 293-297.
- Dirks, M. A., De Los Reyes, A., Briggs-Gowan, M., Cella, D., & Wakschlag, L. S. (2012). Annual Research Review: Embracing not erasing contextual variability in children’s behavior – theory and utility in the selection and use of methods and informants in developmental psychopathology. *Journal Of Child Psychology And Psychiatry, 53*, 558-574. doi:10.1111/j.1469-7610.2012.02537.
- Elliott, S. N., & Gresham, F. M. (2007). *Social Skills improvement System (SSIS) – Performance Screening Guide*. Minneapolis, MN: Pearson Assessment.
- Elliott, S. N., & Gresham, F. M. (2007a). *SSIS classwide intervention program teacher’s guide*. Minneapolis, MN: NCS Pearson.
- Elliott, S. N., Busse, R. T., & Gresham, F. M. (1993). Behavior rating scales: Issues of use and development. *School Psychology Review, 22*, 313-321.



- Fabiano, G. A., Vujnovic, R., Naylor, J., Parsieau, M., & Robins, M. (2009). An investigation of the technical adequacy of a daily behavior report card for monitoring progress of students with ADHD in special education placements (2009). *Assessment for Effective Intervention, 24*, 231-241.
- Feeley, T. H. (2002). Comment on halo effects in rating and evaluation research. *Human Communication Research, 28*, 578-586.
- Floyd, R. G., & Bose, J. E. (2003). Behavior rating scales for assessment of emotional disturbance: A critical review of measurement characteristics. *Journal of Psychoeducational Assessment, 21*, 43-78.
- Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*, 188-193.
- Gadow, K. D. (1986). *Peer Conflict Scale*. Stony Brook: State University of New York: Department of Psychiatry.
- Gadow, K. D. & Sprafkin, J. (2002). *Childhood Symptom Inventory-4 screening and norms manual*. Stony Brook: Checkmate Plus.
- Gadow, K. D., & Sprafkin, J. (2008). ADHD Symptom Checklist-4 2008 Manual. *Stony Brook, NY: Checkmate Plus*.
- Gadow, K. D., Nolan, E. E., Paolicelli, L. M., & Sprafkin, J. (1991). A procedure for assessing the effects of methylphenidate on hyperactive children in public school settings. *Journal of Clinical Child Psychology, 20*, 268-276.
- Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. W. H. Freeman.
- Golfried, M. R., & Kent, R. N. (1972). Traditional versus behavioral personality assessment: A comparison of methodological and theoretical assumptions. *Psychological Bulletin, 77*, 409.
- Gresham, F. M. (2004). Current status and future directions of school-based behavioral interventions. *School Psychology Review, 33*, 326-334.
- Gresham, F. M. (2008). Best practices in diagnosis in a multi-tier problem solving approach. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (5th ed., vol. 2, pp. 281-294). Bethesda, MD: National Association of School Psychologists.

- Gresham, F. M. (2011). Social behavioral assessment and intervention: Observations and impressions. *School Psychology Review, 40*, 275- 283.
- Gresham, F. M., & Elliott, S. N. (1990). *The Social Skills Rating System*. Circle Pines, MN: American Guidance Service.
- Gresham, F. M., & Elliott, S.N. (2008). *Social Skills Improvement System: Rating Scales Manual*. Minneapolis, MN: Pearson Assessments.
- Gresham, F. M., Carey, M. P. (1988). Research methodology and measurement. In Witt, J., Elliott, S., & F. Gresham (Eds). *Handbook of behavior therapy in education* (pp. 37-65). New York, NY, US: Plenum Press.
- Gureasko-Moore, D. P., DuPaul, G. J., & Power, T. J. (2005). Stimulant treatment for Attention-Deficit/Hyperactivity Disorder: Medication monitoring practices of school psychologists. *School Psychology Review, 34*, 232-245.
- Harrison, S.E., Jaffery, R., Johnson, A., Chafouleas, S.M., Riley-Tillman, T.C., & Christ, T.J. (2011, February). *Evaluating the effectiveness of a Direct Behavior Rating training module*. Poster presentation at the National Association of School Psychologists Annual Convention, San Francisco, CA.
- Hartmann, D. P., Roper, B. L., & Bradford, D. C. (1979). Some relationships between behavioral and traditional assessment. *Journal of Behavioral Assessment, 1*, 3-21.
- Harwell, M. (1999). Evaluating the validity of educational rating data. *Educational And Psychological Measurement, 59*, 25-37. doi:10.1177/00131649921969712.
- Hersh, R., & Walker, H. M. (1983). Great expectations: Making schools effective for all students. *Policy Studies Review, 2*(1), 147-188.
- Hintze J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and settings: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review, 33*, 258-270.
- Hintze, J. M. (2005). The psychometrics of direct observation. *School Psychology Review, 34*, 507-519.
- Hintze, J. M., Volpe, R. J., & Shapiro, E. S. (2008). Best practices in systematic direct observation of student behavior. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (Vol. 2, pp. 319-335). Bethesda, MD: National Association of School Psychologists.

- Hoagwood, K., & Erwin, H. D. (1997). Effectiveness of school-based mental health services for children: A 10-year research review. *Journal of Child and Family Studies*, 6, 435-451.
- Hyman, Wojtowicz, A., Duk Lee, K., Haffner, M. E., Fiorello, C. A., Storlazzi, J. J. & Rosenfeld, J (1998). School-based methylphenidate placebo protocols: Methodological and practical issues. *Journal Of Learning Disabilities*, 31, 581-594.
- Individuals with Disabilities Education Act (IDEA). (2004). 20 U.S.C. Section 1400 (1990).
- Jacobson N. S. & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59, 12-19.
- Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (2007). Response to intervention at school: The science and practice of assessment and intervention. In S. Jimerson, M. Burns, & A. VanDerHeyden (Eds.), *Handbook of response to intervention: The science and practice of assessment and intervention* (pp. 3-9). New York, NY: Springer Science and Business Media, LLC.
- Jones, E. E., & Nisbett, R. E. (1971). *The actor and the observer: Divergent perceptions of the causes of behavior* (p. 16). New York: General Learning Press.
- Johnston, J. M., & Pennypecker, H.S. (1980). *Strategies and tactics of human behavioral research*. Hillsdale, NJ: Erlbaum.
- Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement*, 6, 125-160.
- Kazdin, A. E. (1979). Situational specificity: The two-edged sword of behavioral assessment. *Behavioral Assessment*, 1, 57-75.
- Kovacs, M. (2011). *Children's Depression Inventory 2* (2<sup>nd</sup> edition). North Tonawanda: NY: Multi-Health Systems Inc.
- Kraemer, H., Measelle, J., Ablow, J., Essex, M., Boyce, W., & Kupfer, D. (2003). A new approach for integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry*, 160, 1566-1577.

- Kratochwill, T. R., & Bergan, J. R. (1990). *Behavioral consultation in applied settings: An individual guide*. New York: Plenum.
- Kupersmidt, J., Coie, J., & Dodge, K. (1990). The role of peer relationships in the development of disorder. In S. Asher & J. Coie (Eds.), *Peer rejection in childhood* (pp. 274-308). New York: Cambridge University Press.
- Lahey, B.B., Gendrich, J.G., Gendrich, S.I., Schnelle, J.F., Gant, D.S., & McNees, M.P. (1977). An evaluation of daily behavior report cards with minimal teacher and parent contacts as an efficient method of classroom intervention. *Behavior Modification, 1*, 381-394.
- LeBel, T.J., Kilgus, S.P., Briesch, A.M., & Chafouleas, S.M. (2010). The impact of training on the accuracy of teacher-completed Direct Behavior Ratings (DBRs). *Journal of Positive Behavioral Interventions, 12*, 55-63.
- Loney, J. & Milich, R. (1982). Hyperactivity, inattention, and aggression in clinical practice. In M. Wolraich & D. K Routh (Eds.), *Advancements in developmental and behavioral pediatrics, Vol. 3*. (pp.113-147). Greenwich, CT: JAI.
- McWilliam, R. A., & Ware, W. B. (1994). The reliability of observations of young children's engagement: An application of generalizability theory. *Journal of Early Intervention, 18*, 34-47.
- Meier, S. T., McDougal, J. L., & Bardos, A. (2008). Development of a change-sensitive outcome measure for children receiving counseling. *Canadian Journal Of School Psychology, 23*, 148-160. doi:10.1177/0829573507307693
- Merrell, K. W. (2003). *Behavioral, social and emotional assessment of children and adolescents* (2<sup>nd</sup> ed.) Mahwah, NJ: Erlbaum.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749. doi:10.1037/0003-066X.50.9.741
- Nunnally, J., & Bernstein, I. (1994). *Psychometric Theory*. (3<sup>rd</sup> ed.). New York: McGraw Hill.
- Nguyen, N., Whittlesey, S., Scimeca, K., DiGiacomo, D., Bui, B., Parsons, O., & ... Paddock, D. (1994). Parent-child agreement in prepubertal depression: Findings with a modified assessment method. *Journal Of The American Academy Of Child & Adolescent Psychiatry, 33*, 1275-1283. doi:10.1097/00004583-199411000-00008.

- Parkes, J. (2000). Relationship between reliability and cost of performance assessment. *Education Policy Analysis Archives*, 8, 16. Retrieved from <http://epaa.asu.edu/ojs/article/view/407/530>.
- Pelham, W. r., Fabiano, G. A., & Massetti, G. M. (2005). Evidence-based assessment of attention deficit hyperactivity disorder in children and adolescents. *Journal Of Clinical Child And Adolescent Psychology*, 34, 449-476. doi:10.1207/s15374424jccp3403\_5.
- Pelham, W. E., Hoza, B., Pillow, D. R., Gnagy, E. M., Kipp, H. L., Greiner, A. R., et al. (2002). Effects of methylphenidate and expectancy on children with ADHD: Behavior, academic performance, and attributions in a summer treatment program and regular classroom settings. *Journal of Consulting and Clinical Psychology*, 70, 320-325.
- Power, T. J., DuPaul, G. J., Shapiro, E. S., & Kazak, A. E. (2003). *Promoting children's health: Integrating school, family, and community*. New York, NY US: Guilford Press.
- Reynolds, C., & Kamphaus, R. (2004). *Behavioral Assessment System for Children-Second Edition*. Minneapolis, MN: Pearson Assessments.
- Riley-Tillman, T., Chafouleas, S. M., & Briesch, A. M. (2007). A school practitioner's guide to using daily behavior report cards to monitor student behavior. *Psychology in the Schools*, 44, 77-89.
- Riley-Tillman, T., Chafouleas, S. M., & Briesch, A. M., & Eckert, A. (2008). Daily behavior report cards and systematic direct observation: An investigation of the acceptability, reported training and use, and decision reliability among school psychologists. *Journal Of Behavioral Education*, 17, 313-327.
- Riley-Tillman, T., Chafouleas, S. M., Christ, T., Briesch, A. M., & LeBel, T. J. (2009). The impact of item wording and behavioral specificity on the accuracy of direct behavior ratings (DBRs). *School Psychology Quarterly*, 24, 1-12.
- Riley-Tillman, T.C., Chafouleas, S.M., Sassu, K.A., Chanese, J.A.M., & Glazer, A.D. (2008). Examining the agreement of Direct Behavior Ratings and Systematic Direct Observation for on-task and disruptive behavior. *Journal of Positive Behavior Interventions*, 10, 136-143. doi:10.1177/1098300707312542.
- Riley-Tillman, T.C., Christ, T.J., Chafouleas, S.M., Boice, C.H. & Briesch, A.M. (2010). The impact of observation duration on the accuracy of data obtained from Direct Behavior Rating (DBR). *Journal of Positive Behavior Interventions*, 13, 119-128.

- Salvia, J., Ysseldyke, J. E., & Bolt, S. (2004). *Assessment in special and inclusive education*. Boston, MA: Houghton Mifflin.
- Saudargas, R. A., & Zanolli, K. (1990). Momentary time sampling as an estimate of percentage time: A field validation. *Journal of Applied Behavior Analysis, 23*, 533.
- Schlientz, M.D., Riley-Tillman, T.C., Briesch, A.M., Walcott, C.M., & Chafouleas, S.M. (2009). The impact of training on the accuracy of Direct Behavior Ratings (DBRs). *School Psychology Quarterly, 24*, 73-83.
- Shapiro, E. S. (1988). Behavioral assessment. In J. C. Witt, S. N. Elliott, F. M. Gresham (Eds.), *Handbook of behavior therapy in education* (pp. 67-98). New York, NY US: Plenum Press.
- Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal Of Mathematical And Statistical Psychology, 34*, 133-166.
- Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage Publications: Newbury Park: CA.
- Shinn, M. R. (2002). Best practices in using curriculum-based measurement in a problem-solving model. *Best practices in school psychology IV, 1*, 671-697.
- Shinn, M. R., Tindal, G. A., Spira, D. A., & Marston, D. (1987). Practice of learning disabilities as social policy. *Learning Disability Quarterly, 10*, 17-28. doi:10.2307/1510751.
- Smith, P. L. (1981). Gaining accuracy in generalizability theory: Using multiple designs. *Journal Of Educational Measurement, 18*, 147-154.
- Spool, M. D. (1978). Training programs for observers of behavior: A review. *Personnel Psychology, 31*, 853-888.
- Suen, H. K., & Pui-Wa, L. (2007). Classical versus generalizability theory of measurement. *Educational Measurement, 4*, 3-14.
- Swiss Society for Research in Education Working Group. (2010). EduG (Version 6.1). Retrieved from <http://www.irdp.ch/edumetrie/englishprogram.html>

- Tilly, W. D. (2008). The evolution of school psychology to science-based practice: Problem solving and the three-tiered model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 17–36). Bethesda, MD: National Association of School Psychologists.
- Traub, R. E. (2005). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*, 8-14.
- Tversky, B., & Marsh, E. J. (2000). Biased retellings of events yield biased memories. *Cognitive Psychology, 40*, 1-38.
- Volpe, R. J., & Briesch, A. M. (2012). Generalizability and dependability of single-item and multiple-item direct behavior rating scales for engagement and disruptive behavior. *School Psychology Review, 41*, 246-261.
- Volpe, R. J., Briesch, A. M., & Gadow, K. D. (2011). The efficiency of behavior rating scales to assess inattentive-overactive and oppositional-defiant behaviors: Applying generalizability theory to streamline assessment. *Journal Of School Psychology, 49*, 131-155.
- Volpe, R. J. & Chafouleas, S. M. (2011). Assessment of externalizing behavioral deficits. In M. Bray T. J. Kehle (Eds.), *The Oxford handbook of school psychology*. (pp. 284-311). New York: Oxford Press.
- Volpe, R. J., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in classroom settings: A review of seven coding schemes. *School Psychology Review, 34*, 454-474.
- Volpe, R. J., & Gadow, K. D. (2010). Creating abbreviated rating scales to monitor classroom inattention-over activity, aggression, and peer conflict: reliability, validity and treatment sensitivity. *School Psychology Review, 39*, 350-363.
- Volpe, R. J., & Gadow, K. D., Blom-Hoffman, J., & Feinberg, A. B. (2009). Factor-analytic and individualized approaches to constructing brief measures of ADHD behaviors. *Journal of Emotional and Behavioral Disorders, 17*, 118-128.
- Volpe, R. J., McConaughy, S. H., & Hintze, J. M. (2009). Generalizability of classroom behavior problem and on-task scores for the direct observation form. *School Psychology Review, 38*, 382-401.



- Volpe, R. M., & Briesch, A.M. (2012). Generalizability and dependability of single-item and multiple-item direct behavior rating scales for engagement and disruptive behavior. *School Psychology Review, 41*, 246-261.
- Walker H., Irvin L., Noell, J.,& Singer G. (1992). A construct score approach to the assessment of social competence: Rationale, technological considerations, and anticipated outcomes. *Behavior Modification, 16*, 448-474.
- Walker, H. M., & Severson, H. H. (1990). *Systematic screening for behavior disorders (SSBD): User's guide and technical manual*. Longmont, CO: Sopris West.
- Walker, H. M., Seeley, J. R., Small, F., Severson, H. H., Graham, B. A., Feil, E. G., Serna, L., ... Forness, S. R. (2009). A randomized control trial of the First Steps to Success early intervention: Demonstration of program efficacy outcomes in a diverse urban school district. *Journal of Emotional and Behavioral Disorders, 17*, 197-212.
- Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. *Handbook of Statistics, 26*, 1-44.
- Wentzel, K. R. (1993). Does being good make the grade? Relations between academic and social competence in early adolescence. *Journal of Educational Psychology, 85*, 357-364.
- Wilson, M. S., & Reschly, D. J. (1996). Assessment in school psychology training and practice. *School Psychology Review, 25*, 9-23.
- Wolraich, M. L., Feurer, I., Hannah, J. N., Pinnock, T. Y., & Baumgaertel, A. (1998). Obtaining systematic teacher reports of disruptive behavior disorders utilizing DSM-IV. *Journal of Abnormal Child Psychology, 26*, 141– 152.



# APPENDIX: INSTITUTIONAL REVIEW BOARD APPROVAL SHEET

## ACTION ON PROTOCOL APPROVAL REQUEST



Institutional Review Board  
Dr. Robert Mathews, Chair  
131 David Boyd Hall  
Baton Rouge, LA 70803  
P: 225.578.8692  
F: 225.578.6792  
[irb@lsu.edu](mailto:irb@lsu.edu) | [lsu.edu/irb](http://lsu.edu/irb)

**TO:** Frank Gresham  
Psychology

**FROM:** Robert C. Mathews  
Chair, Institutional Review Board

**DATE:** January 29, 2013  
**RE:** IRB# 3349

**TITLE:** Generalizability and Dependability of Brief Behavior Rating Scales for Social Skills

**New Protocol/Modification/Continuation:** New Protocol

**Review type:** Full  Expedited  **Review date:** 1/29/2013

**Risk Factor:** Minimal  Uncertain  Greater Than Minimal

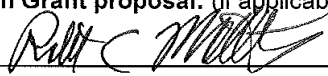
**Approved**  **Disapproved**

**Approval Date:** 1/29/2013 **Approval Expiration Date:** 1/28/2014

**Re-review frequency:** (annual unless otherwise stated)

**Number of subjects approved:** 6

**Protocol Matches Scope of Work in Grant proposal:** (if applicable)

**By:** Robert C. Mathews, Chairman 

**PRINCIPAL INVESTIGATOR: PLEASE READ THE FOLLOWING –**  
**Continuing approval is CONDITIONAL on:**

1. Adherence to the approved protocol, familiarity with, and adherence to the ethical standards of the Belmont Report, and LSU's Assurance of Compliance with DHHS regulations for the protection of human subjects\*
2. Prior approval of a change in protocol, including revision of the consent documents or an increase in the number of subjects over that approved.
3. Obtaining renewed approval (or submittal of a termination report), prior to the approval expiration date, upon request by the IRB office (irrespective of when the project actually begins); notification of project termination.
4. Retention of documentation of informed consent and study records for at least 3 years after the study ends.
5. Continuing attention to the physical and psychological well-being and informed consent of the individual participants, including notification of new information that might affect consent.
6. A prompt report to the IRB of any adverse event affecting a participant potentially arising from the study.
7. Notification of the IRB of a serious compliance failure.

**8. SPECIAL NOTE:**

*\*All investigators and support staff have access to copies of the Belmont Report, LSU's Assurance with DHHS, DHHS (45 CFR 46) and FDA regulations governing use of human subjects, and other relevant documents in print in this office or on our World Wide Web site at <http://www.lsu.edu/irb>*

## VITA

Lisa R. Libster Minor was born in Silver Spring, Maryland. Lisa graduated from Sherwood High School in 2000. Lisa attended Towson University and graduated Summa Cum Laude with a Bachelor of Science degree in psychology. Lisa also received departmental honors in psychology, clinical option, and is a graduate of the Honors College. In 2006, Lisa began her studies in the school psychology doctoral program at Louisiana State University, under the guidance of Dr. Frank Gresham, and earned a Masters of Arts degree in 2009. In 2012, Lisa completed an APA accredited internship in clinical psychology with a focus in Applied Behavior Analysis at the May Institute under the supervision of Dr. Gary Pace and Dr. Melanie DuBard. She also became credentialed as a Board Certified Behavior Analyst. In 2013, Lisa began a post-doctoral fellowship at the Christian Sarkine Autism Treatment Center at the Indiana University School of Medicine. Lisa's current research and clinical interests include social skills deficits, behavioral intervention, Autism Spectrum Disorders, parent training, response-to-intervention and classroom management.