The Pennsylvania State University

The Graduate School

Department of Educational Psychology, Counseling, and Special Education

GENERALIZABILITY OF DIRECT BEHAVIOR RATINGS BY TRAINED AND UNTRAINED RATERS

A Dissertation in

School Psychology

by

Bradley T. Leposa

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2017

The dissertation of Bradley T. Leposa was reviewed and approved* by the following:

Peter M. Nelson
Assistant Professor of School Psychology
Dissertation Advisor
Chair of Committee

Barbara A. Schaefer
Associate Professor of Education
Director of Training/Professor in Charge for Graduate Programs in School Psychology

Richard J. Hazler
Professor of Education (Counselor Education)

Mark T. Greenberg
Professor of Human Development and Psychology
Edna Peterson Bennett Endowed Chair in Prevention Research

*Signatures are on file in the Graduate School.

ABSTRACT

Direct Behavior Ratings (DBRs) are an emerging tool for gathering data to use in decisions within both traditional and expanded roles for school psychologists. The present study contains two components examining DBRs. The first component uses Generalizability Theory (GT) to examine agreement between two raters using DBRs in two different classrooms, one with trained raters and the other with untrained raters. The second component uses a focus group to solicit raters' views regarding DBRs, raters' views regarding a DBR training module, and the reasoning behind raters' ratings. The G studies provide evidence that rater load (i.e., the number of students rated at one time) may impact rater disagreement; however, training had no discernable effect on reliability. Results from the focus group indicate raters found the DBRs mechanically easy to fill out, but that they had difficulty rating and differentiating between Respectful and Disruptive Behavior. For school psychologists in practice, the results of the present research suggest characteristics of the sample of students being rated (i.e., number of students) should be considered when interpreting DBR scores. For further development of DBR items, additional constructs suited to educators' needs, such as a construct targeting social skills, may be beneficial.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

First, I would like to thank Dr. Nelson, my advisor and committee chair, for his advice and encouragement during the process of completing my dissertation. I also would like to thank my committee members, Dr. Schaefer, Dr. Hazler, and Dr. Greenberg, for agreeing to be a part of my committee and for their invaluable input and advice. In addition, I want to thank Dr. Suen for his input and advice regarding the design of my study. Lastly, I want to thank Dr. Diibor and Dr. Olley of Baltimore City Public Schools for their help and support in getting my study through the IRB process of Baltimore City.

In addition to the Penn State faculty and Baltimore officials who helped me, I also need to thank my parents for their help and support throughout graduate school. I would not have made it as far as I have without them. I need to thank the members of my cohort for their support throughout our time in graduate school as well. Lastly, I must thank my friends who helped me relax and take my mind, briefly, off my studies.

Chapter 1: Introduction

The prevalence of emotional and behavioral disorders among children and youth is an important national issue (Federal Interagency Forum on Child and Family Statistics, 2013). Because institutions for providing mental health services to children and youth are fragmented and difficult to access, public officials, researchers, and mental health professionals recognize schools as potential venues for preventing, identifying, and treating mental illness (Hoagwood & Erwin, 1997). Within schools, school psychologists serve as mental health professionals who assist students with mental health problems. In their traditional role, school psychologists assess students with social, emotional, and behavioral problems to determine eligibility for special services. However, several interrelated initiatives exist to expand the role of school psychologists to better promote students' mental health (Merrell, 2002; Merrell, Ervin, & Peacock, 2012). More specifically, professionals within the field of school psychology continue to advocate for a multi-tiered prevention model of service, data-based problem solving in schools, and the dissemination of evidence-based assessments and interventions (Kratochwill, 2007; Strein, Hoagwood, & Cohn, 2003; Tilly, 2008). Because of the unique data collection requirements associated with those new roles, a continuing need exists to develop new data-collection tools suited to the unique demands of those efforts and programs (Chafouleas, Volpe, Gresham & Cook, 2010; Lewis, Barbara, Mitchell, Bruntmeyer, & Sugai, 2016).

Data Collection in Schools

Data collection is an important component of all roles of a school psychologist. In the traditional role of school psychology practice, school psychologists use data collected during assessments to inform high-stakes classification and eligibility decisions (Jacob, Decker, & Hartshorne, 2011; Merrell et al., 2012). Within multi-tiered systems of support, school

psychologists may collect data to help (a) screen students for risk, (b) progress monitor students to determine tier placement, (c) create a working hypothesis regarding the cause of a problem, (d) select or design an intervention to address a problem, and (e) evaluate the effects of an intervention (Tilly, 2008). In the effort to disseminate evidence-based practices and interventions, school psychologists help collect data to evaluate the efficacy of an intervention or practice in the setting where the service is offered (Merrell et al., 2012).

Utility of Existing Data Collection Tools in Traditional and Emerging Roles

The directness and structure of a given data-collection tool affects its utility for different data-collection purposes (Chafouleas, 2011; Chafouleas et al., 2010; Christ, Riley-Tillman, & Chafouleas, 2009; Merrell, 2011). Directness refers to how close the collected data are to the targeted construct or behavior. Structure refers to the flexibility of a tool to be modified to suit different data-collection purposes and to allow respondents to generate their own responses. For directness, data-collection tools fall into the dichotomous categories of direct and indirect, but for structure, data-collection tools fall along a continuum (Christ et al., 2009; Merrell, 2011).

Indirect data-collection tools are useful for diagnostic purposes in the traditional role of school psychology because they often yield psychometrically defensible data suitable for decisions requiring comparisons with other students or a criterion (Christ et al., 2009). However, indirect tools often are too resource intensive to be feasible to collect data for decisions regarding screening and tier placement in multi-tiered systems of support. In addition, indirect tools sometimes lack utility in problem solving because they cannot convey information about the context in which the data were collected (Chafouleas, 2011). In contrast, direct tools can convey information about the context where data were collected. Direct tools also

do not rely on the memories of informants, which avoids potential biases and inaccuracies associated with memory. However, direct tools often are too resource intensive to be feasible for efficient use in screening, progress monitoring, and problem-solving in multi-tiered systems of support (Chafouleas, 2011; Chafouleas et al., 2010). Moreover, unstructured direct tools do not yield numerical data that quantify behavior, which also is needed for utility in screening, progress monitoring, and problem solving (Christ et al., 2009). Structured direct tools, such as systemic direct observations (SDOs), can yield data suitable for progress monitoring and problem solving, but SDOs may be too resource intensive in terms of time and training to be feasible for widespread use in those roles (Chafouleas, 2011; Christ et al., 2009).

Direct Behavior Ratings (DBRs)

The term direct behavior ratings refers to a class of assessment tools with three core defining features: the ratings occur in immediate proximity to the behavior of interest, the rater has firsthand experience of the student being rated, and minimal inference is required to identify the target behavior or behaviors (Chafouleas, Riley-Tillman, & Christ, 2009). DBR researchers conceptualized DBRs as combining the directness of direct assessment tools, such as Systemic Direct Observations (SDOs), with the structure of indirect assessment tools, such as behavior rating scales (Christ et al., (2009). Although the term DBR does not refer to a specific scale type, research on the development of DBR scales converged on the creation of items that consist of a horizontal line along which a rater makes a mark to indicate the frequency, intensity, and duration of the behavioral construct of interest (Chafouleas, 2011). The resulting DBR scales consist of either a single DBR item (DBR-SIS) or multiple DBR items (DBR-MIS) (Chafouleas et al., 2009). The present study focuses on DBR-SISs because DBR-SISs have been the focus of most prior research.

DBRs possess properties that make them useful for collecting data within emerging roles for school psychologists, in particular within multi-tiered systems of support. The directness of DBRs allows them to convey contextual information about data, and the structure of DBRs allows them to yield quantitative data (Chafouleas, 2011). Further, the simplicity of DBRs makes them feasible for teachers and other school staff to use with little or no training and efficient enough to use to collect data multiple times over long time periods. As a result, DBRs are beginning to be used to collect data for screening, problem solving, and progress monitoring within multi-tiered systems of support (Chafouleas, Kilgus, & Hernandez, 2009; Christ et al., 2009).

DBRs are completed by teachers and other school staff who are able to directly observe the student. For progress monitoring and problem solving, DBR data is interpreted for decisions regarding the selection and design of interventions, as well as whether interventions should be continued or discontinued. For screening, DBR data is interpreted for decisions regarding whether students will receive additional supports (Kilgus, 2013). However, research to establish the psychometric quality—or defensibility—of DBRs in these roles is an ongoing process (Christ et al., 2009; Johnson, Miller, Chafouleas, Welsh, Riley-Tillman, & Fabino, 2016).

Reliability of DBRs

Reliability is an important component of the defensibility of any data-collection tool, and evidence for the reliability of a tool is required to establish that a tool is defensible for its intended use (AERA, 2014; IDEA, 2004). Both the ethical codes of the American Psychological Association (APA) and the National Association of School Psychologists (NASP) reflect this requirement because they mandate that data-collection tools have research supporting reliability for their intended uses (APA, 2010; NASP, 2010). The Individuals with Disabilities Education Act of 2004 also reflects

this requirement because it mandates that tools used in assessments of eligibility for special services be reliable (IDEA, 2004).

Because of the importance of reliability in establishing the defensibility of a data-collection tool, DBR researchers used generalizability theory (GT) in several studies to investigate the reliability of DBRs in both naturalistic and controlled environments. Raters are a large source of unreliability in scores derived from DBR-SISs in these G studies (Briesch, Chafouleas, & Riley-Tillman, 2010; Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese, 2007; Christ, Riley-Tillman, Chafouleas, & Boice, 2010; Volpe & Briesch, 2012). The results of GT research with DBRs suggests that DBR-SIS data might only be interpretable within raters rather than across raters (Briesch et al., 2010; Christ et al., 2010). Such a limitation undermines the validity of decisions based on DBR ratings for purposes requiring a comparison to an absolute criterion or across raters, such as for screening. As a result, researchers are interested in the potential of training for DBR users to reduce inconsistencies in scores attributable to raters. However, existing studies examining the psychometric properties of DBRs completed by trained raters focus on DBR accuracy—a slightly different concept than reliability (Chafouleas, Kilgus, Riley-Tillman, Jaffery, & Harrison, 2012; Chafouleas, Riley-Tillman, Jaffery, Miller, & Harrison, 2014; Harrison, Riley-Tillman, & Chafouleas, 2014; Lebel, Kilgus, Briesch, & Chafouleas, 2010). Accuracy concerns the degree to which data collected by a tool reflect the construct of interest and the construct related validity of the tool. In contrast, the reliability of a tool concerns the consistency of scores derived from the tool, such as the consistency of scores between raters (Hintze, 2005). Because of the lack of studies investigating the reliability of DBRs from trained raters, studies that use GT to investigate the reliability of DBRs completed by raters after they have received training are needed.

Rationale and Significance

To address the gap in the literature concerning the reliability of DBRs by trained raters, the main component of the present study uses GT to estimate the amount of variance attributable to raters who received an online training module in the use of DBR. In addition, because previous research with DBRs (i.e., Briesch et al. 2010) highlighted rater fatigue as a potential factor in decreasing reliability, the present study replicates that work with reduced rater demands. In doing so, the present study expands upon previous research to evaluate the generalizability of DBR scores under more favorable conditions (i.e., reduced rater demands and access to training).

Large amounts of variance attributable to raters in the present study would indicate that DBRs, even by trained raters, might not be defensible in some of the data-collection roles researchers propose for them, such as screening and progress monitoring in multi-tiered systems of support. In addition to the G studies, the present study also contains a focus group component to gather information about raters' views regarding DBRs. That information is useful as it may highlight possible reasons for interrater variance on DBRs, as well as future directions for the design and content of DBR items.

Purpose of the Present Study

The present study contains two components. The main component uses GT to estimate the variance attributable to raters in two separate G studies conducted in the naturalistic setting of a school: a G study in which raters were exposed to training and a G study in which raters were not exposed to training. Following the design of the Briesch et al. (2010) generalizability study, the ratings in each G study were completed by two school-based professionals in the same kindergarten classroom over a period of 10 days but with less students rated to reduce rater demands. More specifically, the

first component of the present study addressed the following three related questions:

1. How much variance was attributable to raters in the G study with trained raters compared to previous G studies examining DBRs by untrained raters in naturalistic settings?
2. How much variance was attributable to raters in the current G study with untrained raters compared to Briesch et al. (2010)?
3. How much variance was attributable to raters in the G study with trained raters compared to the G study with untrained raters?

In addition to the G studies, the present study also contains a focus group component. Focus groups can help facilitate interpretation of data collected by quantitative studies (Liamputtong, 2011). In particular, focus groups can gather information about how individuals arrive at conclusions they draw (Stewart, Shamdasani, & Rock, 2007). As a result, the researcher conducted a focus group at the conclusion of the G studies to obtain information about the reasoning behind raters' ratings, as well as raters' views of DBRs in terms of feasibility and utility. The information collected during the focus group will help (a) suggest reasons for the large amount of interrater variance observed in the G studies of DBRs, (b) inform future efforts to enhance the feasibility of DBRs, (c) inform future efforts to design the wording of DBR items, and (d) inform future efforts to develop behavioral constructs measured by DBRs.

Chapter 2: Literature Review

Direct behavior ratings (DBRs) are a data-collection tool developed for uses within both traditional and emerging roles for school psychologists. The present study concerns the generalizability of DBRs completed by trained raters in a school. To provide an appropriate context for the current research questions, a brief discussion of the mental health problems facing students is followed by a discussion of the traditional and emerging roles of school psychologists. The tools school psychologists use to collect and interpret data are discussed briefly before an in-depth review of the psychometric issues related to DBRs.

Mental Health Needs of Children and Adolescents

The prevalence of emotional and behavioral disorders among children and youth is an area of national priority and concern (FIFCFS, 2013; United States Department of Health and Human Services, 1999; United States Public Health Service, 2000). In the United States, one in five children and adolescents experience signs and symptoms of psychopathology within a year, and 5% of U.S. children and adolescents experience extreme functional impairment resulting from mental illness (FIFCFS, 2013; USDHHS, 1999). Further, emotional and behavioral disorders among children and youth are of particular concern because mental health is a critical component in children's learning and general health (USPHS, 2000). For example, children and adolescents with serious emotional and behavioral problems are at increased risk for school failure and other negative developmental outcomes (New Freedom Commission on Mental Health, 2003).

Despite the prevalence of emotional and behavioral disorders among children and youth and the resulting negative developmental outcomes, systems for providing mental health services to children and youth remain fragmented and difficult to access (NAMHC, 2001; NFCMH, 2003; USDHHS, 1999). As a result, public officials, mental health professionals, and researchers in the

field of mental health recognize schools as potentially important venues for providing mental health services to children and adolescents because all children and adolescents must attend these institutions (Hoagwood & Erwin, 1997; USDHHS, 1999; USPHS, 2000). Roles in addressing mental health issues among children and youth suggested by researchers, public officials, and mental health professionals for schools include: (a) early detection of emotional and behavioral problems before they become severe, (b) assessment and identification of emotional and behavioral disorders, (c) prevention of emotional and behavioral disorders, and (d) intervention to address emotional and behavioral disorders (NFCMH, 2003; USDHHS, 1999; USPHS, 2000).

Roles of School Psychology

School psychologists work within schools to promote the mental health of children and youth in several ways. In their traditional role, school psychologists assess students referred for emotional and behavioral difficulties to determine eligibility for special services under the Individuals with Disabilities in Education Act (IDEA) or Section 504 of the Rehabilitation Act of 1973 (Jacob et al., 2011; Merrell et al., 2012). In addition, in response to the numerous mental health issues facing children and adolescents, several initiatives exist to expand the role of school psychologists to have a greater impact on students' social and emotional development (Merrell, 2002; Merrell et al., 2012). These roles include involvement with multi-tiered systems of support, facilitating a data-based and formative approach to problem solving, and increased involvement in efforts to disseminate evidence-based assessments and interventions directed at improving students' mental health (Hoagwood & Johnson, 2003; Kratochwill, 2007; Merrell et al., 2012; Strein et al., 2003; Tilly, 2008).

Traditional role of school psychologists. In the traditional model of school psychology practice, school psychologists address students' emotional and behavioral problems by gathering and interpreting data as part of a

multi-disciplinary team to determine whether students meet criteria for inclusion in legally defined categories. These categories grant students access to special services, and they allow students to receive special accommodations within their schools (Merrell et al., 2012). Two laws defining these categories are the IDEA and Section 504 of the Rehabilitation Act of 1973. The IDEA classifies students with emotional and behavioral disorders in the categories of emotional disturbance (ED) or other health impairment (OHI) (IDEA, 2004; Jacob et al., 2011). Under Section 504, students with emotional and behavioral disorders sometimes qualify for special services and accommodations if the assessment indicates they have a "…mental impairment that substantially limits one or more major life activities…" (Jacob et al., 2011; 29 U.S.C. § 794). Students with emotional and behavioral disorders listed in the Diagnostic and Statistical Manual of Mental Disorders (DSM), such as Major Depression, sometimes meet this definition (Jacob et al., 2011).

Expanded role of school psychologists. School psychologists' traditional roles are limited to gathering and interpreting data for classification and eligibility decisions regarding students with social, emotional, and behavioral problems (Merrell et al., 2012). This practice model limits school psychologists' involvement in designing, selecting, and implementing interventions to promote students' mental health. To further build the capacity of schools to promote students' mental health, school psychologists are part of several interrelated initiatives that expand their role in addressing and preventing students' social, emotional, and behavioral problems (Hoagwood & Johnson, 2003; Kratochwill, 2007; Merrell et al., 2012; Strein et al., 2003; Tilly, 2008).

Two of the major interrelated initiatives involving school psychologists are the implementation of multi-tiered systems of support in schools and the dissemination of evidence-based practices and interventions

(Merrell et al., 2012). Multi-tiered systems of support involve several tiers of interventions that differ in the intensity and frequency of problems addressed and the corresponding supports provided (Jimerson, Burns, & VanderHeyden, 2016). Within multi-tiered systems of support, educational professionals use a process called problem solving to design, select, and evaluate interventions (Tilly, 2008). Evaluating intervention effectiveness through the problem solving process evaluates interventions in their local context (Stoiber & Maribeth, 2016), and evaluating interventions in their local context is a method through which school psychologists help disseminate evidence-based practices and interventions (Merrell et al., 2012).

Multi-tiered systems of support used in schools adopt several elements from tiered prevention models used in the field of public health including screening, data-based decision-making, an emphasis on prevention, and a focus on populations rather than just individuals (Strein et al., 2003). These models contain multiple levels of prevention that differ in the needs of the individuals served and the intensity of the interventions delivered. At the first tier, professionals deliver interventions to an entire population regardless of the level of risk or the presence of mental health problems. These interventions build resiliency and enhance protective factors among the entire school population to prevent problems from occurring. At the second tier, professionals deliver more intensive interventions to small groups of students identified as being at greater risk for developing mental health problems. At the third tier, professionals deliver intensive one-on-one interventions to students who display symptoms or behaviors characteristic of a mental illness (Strein et al., 2003).

Schoolwide Positive Behavior Supports (SWPBS) is one example of a multi-tiered system of support used in schools to manage disruptive behavior among students (Sugai & Horner, 2009). In multi-tiered systems of support used in schools, such as SWPBS, school psychologists help gather and

interpret data through screening and progress monitoring to inform decisions regarding students' placements on tiers and the interventions students receive (Merrell et al., 2012; Sugai & Horner, 2009). Problem solving is an integral component of multi-tiered systems of support in schools because it is an important part of that decision-making process.

Problem solving in multi-tiered systems of support. The problem-solving process helps school psychologists design and select interventions for use in multi-tiered systems of support (Merrell et al., 2012; Sugai & Horner, 2009; Tilly, 2008). In problem solving, school psychologists define problems as a discrepancy between actual and desired performance (Deno, 2002; Tilly, 2008). The process for reducing or eliminating this discrepancy involves four steps (Deno, 2002; Tilly, 2008). In the first step, school psychologists help identify, conceptualize, and define a problem through data collection. In the second step, school psychologists may collect additional data to form a working hypothesis regarding the cause of the problem. In the third step, school psychologists help design or select an intervention that professionals in the school then implement to address the problem. In the fourth step, school psychologists help evaluate the intervention to determine whether it is successful in alleviating the problem (i.e., progress-monitoring). If the evaluation reveals that the intervention is successful, then the intervention is continued. If the evaluation reveals that the intervention is not successful, then the problem-solving process returns to steps one or two and begins again (Deno, 2002; Tilly, 2008). Students who fail to respond to interventions may also move to a higher tier of support with more intense interventions (Merrell et al., 2012; Sugai & Horner, 2009; Tilly, 2008).

Dissemination of evidence-based practices. Evaluating the effectiveness of interventions through problem solving is one part of the effort to ensure that interventions used by school psychologists are evidence-based. The term evidence-based practices refers to service practices, such as assessments or

interventions, that have a body of scientific knowledge demonstrating effectiveness in the setting where the practices are offered (APA, 2006; Kratochwill, 2007). In psychology and education, evidence-based practice refers to the movement to identify, disseminate, and promote the use of practices with demonstrated research support (Kratochwill, 2007).

The evidence base of a practice consists of two dimensions: treatment efficacy and clinical utility. Treatment efficacy refers to the systematic and scientific evaluation of whether a treatment works as intended. Clinical utility refers to the applicability, feasibility, and usefulness of the intervention in the specific setting where it is used (APA, 2006). Evidence for treatment efficacy generally comes from quasi-experimental and experimental studies, with the most weight given to randomized experimental studies with controls (APA, 2006). However, it can be difficult to generalize the results of experimental studies across populations and settings in the social sciences (Cronbach, 1975). As a result, to ensure the clinical utility of a particular practice, such as an intervention, data must be collected locally to evaluate it (Cronbach, 1975; Deno, 2016; Merrell et al., 2012, p. 156-157). Evaluating the effectiveness of interventions using data-driven problem solving is an example of a way to accomplish this goal (Merrell et al., 2012; Stoiber & Maribeth, 2016).

Importance of Data Collection in Traditional and Expanded Roles

Data collection is an important part of all aspects of problem solving and diagnosis. In the traditional role of school psychology, school psychologists gather data during assessments to inform high-stakes decisions regarding a student's eligibility for special services under legally defined classification systems (Merrell et al., 2012). In the expanded roles of school psychology, school psychologists help gather data to inform decisions in multi-tiered models of support and to evaluate the effectiveness of interventions.

As discussed, within multi-tiered systems of support, school psychologists help gather data for problem solving and to screen students for risk of developing social, emotional, and behavioral problems (Herman, Riley-Tillman, & Reinke, 2012; Merrell et al., 2012). Throughout the problem-solving process, school psychologists help collect and interpret data for two purposes. For the first purpose, school psychologists help collect data about the situational context of the problem and the frequency, intensity, and duration of the problem behaviors. School psychologists then help interpret and use that data to form a working hypothesis regarding the cause of the problem (Deno, 2002; Tilly, 2008). For the second purpose, school psychologists help collect and interpret progress-monitoring data to (a) establish a baseline measure of the observable manifestations of the problem, (b) examine the ongoing effects of interventions designed to address the problem, and (c) evaluate the effects of interventions on the problem after the interventions are complete.

To meet the data-collection needs of school psychologists, the development of tools to collect data, particularly for use within multi-tiered systems of support, continues to receive a great deal of attention in school psychology (Chafouleas, 2011; Chafouleas et al., 2010; Christ et al., 2009; Merrell et al., 2012).

Common Data Collection Tools

The characteristics of data-collection tools used by school psychologists vary in terms of structure and directness. The structure of the tools falls along a continuum, but the directness of the tools fall into dichotomous categories: direct and indirect. The directness and indirectness of a tool influences the utility of the tool for different data-collection purposes within both traditional and expanded roles of school psychology (Chafouleas, 2011; Chafouleas et al., 2010; Christ et al., 2009).

Structured data-collection tools often can collect psychometrically sound quantitative data that is useful for high-stakes classification decisions, screening, and progress monitoring (Chafouleas, 2011). Unstructured data-collection tools often cannot collect such data, but such tools can help obtain contextual details and nuances of a problem that structured assessment tools cannot accommodate, which is useful in problem solving (Christ et al., 2009). Direct data-collection tools have the advantage of not relying on the memory of informants, which can bias data collected by indirect data-collection tools, and direct tools often can convey information about the context in which a problem occurs, which is useful for problem solving. However, direct data-collection tools tend to be resource intensive in terms of time and training, which limits the utility of these tools in screening and progress monitoring because those activities require repeated data collection over relatively long periods of time (Chafouleas, 2011; Chafouleas, 2010).

Interviews are an example of a set of indirect data-collection tools because they rely on the memories of informants (Christ et al., 2009). Interviews can be structured, semi-structured, or unstructured (Merrell et al., 2011). In some cases, structured interviews have strong psychometric foundations, which can be useful for collecting defensible data for high-stakes decisions related to classification and eligibility for special services (Christ et al., 2009). However, structured interviews limit the types of responses interviewees can give, which limits the type of information interviewers can collect. Unstructured interviews give the interviewer and interviewee greater freedom to generate their questions and answers than structured and semi-structured interviews. However, unstructured interviews cannot collect defensible quantitative data for high-stakes classification decisions, progress monitoring, or screening (Christ et al., 2009).

Behavior rating scales. Behavior rating scales are a highly structured form of indirect assessment. Rating scales are often used in high-stakes classification decisions because they yield defensible quantitative data that compares students to their peers or a social, emotional, or behavioral criterion (Chafouleas, 2011; Merrell, 2011). Behavioral rating scales consist of a series of items measuring the presence or absence of several behaviors of interest. Two types of behavior rating scales are narrow band and broad band. Narrow-band behavior rating scales assess functioning in just one domain for a single purpose, such as a measure of externalizing behaviors to evaluate Attention Deficit Hyperactivity Disorder. Broad-band behavior rating scales measure behaviors in many domains and include subscales that evaluate both internalizing and externalizing disorders, such as Oppositional Defiant Disorder and Major Depressive Disorder. Behavior rating scales do not require extensive training to complete and self-report forms exist for older children and adolescents (Merrell, 2011; Merrell et al., 2012).

Despite the ability of behavior rating scales to collect defensible data for use in high-stakes decisions, behavior rating scales possess several characteristics that limit their utility in problem solving, progress monitoring, and screening. Because behavior rating scales are indirect tools, they rely on the memories of informants, which can bias the data they collect. In addition, rating scales typically cannot collect contextually relevant details and nuances regarding a problem, which problem solving requires (Chafouleas, 2010). Moreover, rating scales are too resource intensive for use in roles that require repeated data collection, and rating scales are not sensitive to incremental change, two qualities needed for progress monitoring and sometimes screening in multi-tiered systems of support (Chafouleas, 2011).

Direct observations. Direct observations can provide contextually relevant information about a problem because data are obtained at the time and place a behavior occurs. Types of direct observations include narrative observations, antecedent-behavior-consequences (ABC) narratives, and systemic-direct observations (SDOs). In narrative observations, observers record general estimates of the frequency, intensity, and duration of target behaviors over the time period of the observation, as well as relevant details about the situational context of target behaviors (Christ et al., 2009). As a result, narrative observations can be useful for gathering data about the situational context of a problem, but the unstructured nature of narrative observations makes it difficult to use them to collect quantitative for progress monitoring and screening. ABC narratives impose a limited amount of structure on the data they collect, which gives these tools utility for gathering data to generate a working hypothesis. However, little research exists on the psychometric quality of data collected using ABC narratives, and ABC narratives are resource intensive in terms of the time and training required to use them. SDOs use highly standardized rating methods to yield quantitative data concerning students' behaviors and the situational contexts where the behaviors occur, which could be useful for screening and progress monitoring. However, SDOs also are time consuming and require trained raters, which limits their utility in screening and progress monitoring (Chafouleas, 2011; Chafouleas et al., 2010).

Direct Behavior Ratings

Direct Behavior Ratings (DBRs) are a set of data-collection tools receiving interest in school psychology because of their potential utility in roles where existing data-collection tools for behavior have limited utility, such as progress monitoring and screening within multi-tiered systems of support (Chafouleas, 2011; Chafouleas et al., 2010; Christ et al., 2009; Herman et al., 2012). DBR researchers conceptualize DBRs as combining the

directness of SDOs with the structure of behavior rating scales, and DBR researchers consider three key properties to be defining features of DBRs. First, DBRs are completed in immediate proximity to the behavior to interest, such as by a classroom teacher on the behavior of a student in his or her classroom. Second, DBRs are completed by individuals with firsthand experience of the behaviors of the student being rated. Third, DBRs require minimal inference on the part of raters to discern the behavior or behaviors being observed (Chafouleas et al., 2009).

Although the term DBR refers to many different scale types, DBR research converged on items in which raters make a mark along a horizontal line to indicate the proportion of time a student engaged in the behavior of interest during a specified time period (Chafouleas, 2011). The resulting items can be used as single-item scales (DBR-SIS) or multi-item scales (DBR-MIS)(Chafouleas et al., 2009). The present study focuses on DBR-SIS because single-item DBRs have been the focus of most prior research.

Like behavior ratings scales, the DBR-SIS and MIS items are structured because respondents respond to a pre-determined item measuring and defining the targeted behavioral construct. Like SDOs, DBRs are direct, meaning that the ratings occur in the context where the behavior occurs. As a result, DBRs carry information about the context where a behavior occurs, making them useful for problem solving and progress monitoring (Chafouleas et al., 2010; Christ et al., 2009). In addition, DBRs are less resource intensive to use than SDOs, making them more feasible to use for data-collection over long periods of time and across multiple contexts (Chafouleas, 2011; Christ et al., 2009). Also, unlike behavior rating scales but similar to SDOs, DBRs are sensitive to incremental changes in behavior. Overall, DBRs combine several characteristics of behavior rating scales and SDOs into a new data-collection tool that researchers suggest may have several uses in schools

(Chafouleas, 2011; Chafouleas et al., 2010; Christ et al., 2009; Herman et al., 2012).

DBR structure. Similar to pain scales used in medicine, DBRs consist of single or multiple item scales in which a single item consists of a horizontal line along which a rater makes a mark to indicate the frequency, intensity, and duration of a behavioral construct of interest. Single item DBRs (DBR-SIS), the focus of the present study, consist of a single horizontal line measuring the behavioral construct of interest. Multiple item DBRs (DBR-MIS) consist of multiple items used to sample the domain of a behavioral construct (Chafouleas, 2011; Christ et al., 2009). DBR researchers initially focused on the development of DBR-SISs because of the possible advantages they offer in terms of feasibility and simplicity (see Christ et al., 2009). As a result, the present study focuses on DBR-SISs.

Vertical lines divide the horizontal line of a DBR item into 11 equal segments numbered from 0 to 10 from left to right. Additional numerical anchors marked 0 (0%), 5 (5%), and 10 (10%) indicating the percentage of time that behaviors occur appear under the vertical lines. Visual anchors such as a smiling face and a frowning face also appear at the left (frowning face) and right (smiling face) ends of the horizontal line. Short instructions defining the behavioral construct to be measured also are on the DBR forms (Chafouleas, 2011; Christ et al., 2009).

DBR rating procedures. Respondents complete DBRs in the context where a behavior occurs. Individuals who commonly complete DBRs include teachers and parents, as well as the target students themselves when using DBRs as a self-report form. The observation intervals over which respondents complete DBRs are relatively brief, most often between 20 minutes and one hour, and respondents complete the ratings at the end of those intervals (Christ et al., 2009). Respondents can complete the DBRs daily, weekly, bi-weekly, or on multiple occasions within a single day. Because they are repeated, DBR

ratings yield time-series data suitable for visual analysis (Christ et al., 2009). The training required to use DBRs is an area of ongoing study (Chafouleas, 2011), and it is the main focus of the present study.

Interpretation of DBR data. As mentioned, DBRs are completed by individuals, most often teachers, who work in the immediate environment of the student being rated (Chafouleas et al., 2009). When used for problem solving, school psychologists and teachers interpret graphical representations of this data to inform decisions regarding the design or selection of interventions, as well as whether an intervention should be continued or discontinued. When used for screening, DBR data is interpreted to inform decisions regarding whether students will receive additional supports.

Constructs rated by DBRs. Conceptually, school psychologists can use DBRs to collect data on a broad range of targets relevant to students' social, emotional, and behavioral functioning (Christ et al., 2009). For example, school psychologists can customize DBRs to fit the specific needs of the intervention they evaluate (Chafouleas, 2011). However, current research focuses on three broad behavioral constructs (Christ et al., 2009). DBR researchers developed these constructs based on information gathered from a review of literature regarding influences on school success, data gathered from the School Wide Information System (www.swis.org), which is a database maintained by the National Center on Positive Behavior Supports (www.pbis.org), and consultations with experts in the field. Based on the information collected, researchers identified several high incidence desirable and undesirable behaviors. Examples of desirable behaviors included writing, hand-raising, responding to questions, listening to teachers, silent reading, and working. Examples of undesirable behaviors included talking-back, arguing, task-refusal, social rudeness, fighting, hitting, and yelling.

The DBR researchers grouped these behaviors into three constructs: academically engaged, disruptive, and compliant/respectful (Christ et al., 2009). The researchers decided on disruptive behavior as one behavioral grouping because of evidence that such behavior leads to lost teaching time and a large percentage of office discipline referrals. Further, disruptive behavior is easy to observe in a classroom setting. The DBR researchers selected defiant and non-compliant behavior as another behavior grouping because of evidence that defiant and non-compliant behavior overwhelms teachers, disrupts learning, threatens the overall safety of the school population, and negatively affects students' chances of school success. The researchers selected academically engaged behavior as the final behavior grouping because of the need to examine behavior indicative of success in a classroom environment, the link between student engagement levels and academic outcomes, and the need to measure non-disruptive off-task behavior not captured by data sources such as office-discipline referrals (see Riley-Tillman et al., 2009).

DBR item wording. Studies of the influence of DBR item wording on rater accuracy informed the wording of the DBR items used to rate the three behavioral constructs (Riley-Tillman et al., 2009). For global versus example-based definitions of behavior, Riley-Tillman et al. (2009) indicated that global definitions of behavioral constructs, rather than lists of specific behaviors, led to more accurate DBR ratings. As a result, the standardized DBR forms for the three constructs use global definitions of behavioral targets, along with a list of examples of the behavior.

For positive versus negative item wording, studies of DBR item wording yielded mixed results across different behavioral constructs (Chafouleas et al., 2013; Riley-Tillman, et al., 2009). For academic engagement, the researchers found that positive wording (academically engaged) generally resulted in more accurate ratings than negative wording (academically

disengaged). The researchers found less conclusive results for disruptive behavior across the studies. In Riley-Tillman et al. (2009), the researchers concluded that whether the DBR item for disruptive behavior was worded positively or negatively did not appear to make a difference for rating accuracy. However, Chafouleas et al. (2013) concluded DBR ratings were generally more accurate for negative wording (disruptive) than positive wording (non-disruptive). The results for negative versus positive wording for compliance/respect were similarly inconclusive. In Riley-Tillman et al., the ratings for compliance/respect did not meet normality assumptions required for statistical analysis. As a result, the researchers did not perform further analysis on ratings of that item; instead, they concluded that the construct needed revision (Riley-Tillman et al., 2009).

In Chafouleas et al. (2014), although accuracy on a revised item measuring compliant/respectful behavior improved, the impact of negative wording (disrespectful) versus positive wording (respectful) on DBR rating accuracy was mixed. The researchers also noted that although rater accuracy improved on the revised item compared to the item measuring compliance/respect in Riley-Tillman et al. (2009), raters still rated respectful behavior less accurately than the other two behavioral constructs. In addition to testing the impact of negative versus positive item wording on rater accuracy, Chafouleas et al. also surveyed the preferences of participants. In the study, participants preferred positive wording for the compliance/respect item, negative wording for the disruptive item, and positive wording for the academic engagement item.

Regarding the items on the standard DBR forms used in the present study, it should be noted that, as discussed, both Riley-Tillman et al. (2009) and Chafouleas et al. (2014) reported problems with the accuracy of ratings on the item measuring compliant/respectful behavior. Several possible issues might cause problems with this item. First, in their

description of the development of the construct, Riley-Tillman et al. mentioned that compliant behavior was opportunity-bound. The term opportunity-bound referred to the fact that students can only display compliant behavior following an adult request. For behavior to be defined as compliant or respectful following an adult request, Riley-Tillman et al. indicated that a student needs to respond appropriately within 5 seconds, otherwise the behavior should be defined as non-compliant/disrespectful; however, that information is not included on the standard DBR forms. As a result, standards for compliance may vary across raters.

Second, the opportunity bound nature of compliant/respectful behavior as defined in current DBR research might result in sample sizes of behavior that are too small to yield consistent ratings. Third, the definition and examples of respectful/compliant behavior included on the standard DBR forms includes interactions with peers. That definition overlaps somewhat with the example of acting aggressively given in the definition of disruptive behavior. Although instructions on the DBR form note that behaviors can co-occur, the possibility still exists that raters might not consistently group behaviors under either disruptive or respectful behavior. To collect information regarding the behaviors raters group under each construct, the present study includes a focus group where information will be gathered regarding the reasoning behind raters' DBR ratings.

In addition, although not the focus of the present study, it should be noted that researchers currently do not agree on a general outcome measure (GOM) analogous to Oral Reading Fluency in academics to assess students' growth in domains relevant to their social, emotional, and behavioral functioning (Chafouleas et al., 2010; Deno, 2002). Moreover, researchers disagree regarding whether GOMs as conceptualized in multi-tiered prevention models for academics are applicable to behavioral domains (Chafouleas et al., 2010). As a result, despite the three existing behavioral constructs

developed by DBR researchers, much more work is needed to develop GOMs in the social, emotional, and behavioral domains (Chafouleas, 2011).

Defensibility of Direct Behavior Ratings

Any data-collection tool used in school psychology must be defensible in the role it is used (Chafouleas, 2011; Chafouleas et al., 2010; Evans & Owens, 2010). To be defensible, a data-collection tool must have research supporting its technical adequacy. Technical adequacy refers to the reliability and validity of the tool, and, if applicable, the norming sample of the tool (AERA, 2014). The ethical codes of the American Psychological Association (APA) and National Association of School Psychologists (NASP), as well as laws governing eligibility for services as a result of a disability, reflect the importance of the reliability and validity of a data-collection tool. Both the APA and NASP ethical codes mandate that data-collection tools have evidence supporting the reliability and validity of the tool for its intended uses (APA, 2010; NASP, 2010). In addition, the Individuals with Disabilities Education Act (IDEA) of 2004 requires that assessments performed to determine eligibility for special services and accommodations for a disability use reliable and valid data-collection tools (IDEA, 2004).

Reliability

Broadly speaking, reliability refers to the consistencies and inconsistencies of scores derived from a data-collection tool. Theories of reliability involve attempts to quantify these consistencies and inconsistencies (Brennan, 2001, 2010; Cronbach, 2004). Two major theoretical frameworks govern these attempts: classical test theory (CTT) and generalizability theory (GT) (Brennan, 2010). Under both theoretical frameworks, sources of inconsistency are called error. What constitutes error is a matter of definition (Brennan, 2010). In other words, theories of reliability attempt to quantify unwanted variance in a measurement due to inconsistency.

Classical test theory. Spearman (1904) is credited with the first attempt to quantify variance in the measurement of an association between two variables (Brennan, 2010; Brown, 1910; Spearman, 1904). Spearman theorized that the observed scores derived from a measurement are not the true values. Rather, he theorized that observed values are confounded with some sort of error. Spearman distinguished three different potential sources of error: (1) probable error, which decreases with increased sample size, (2) error associated with other variables that distort, constrict, or dilate the relationship between the variables being measured, and (3) what he called accidental error, which is the source of error that became associated with the concept of reliability (Spearman, 1904).

Spearman (1910) described a method to quantify the size of these accidental errors by measuring the size of the discrepancies between successive measures of the same thing (Spearman, 1910). To perform this calculation, Spearman recommended separating a series of measurements into groups so that any differences between the means of the two groups could only be attributed to accidental factors (Spearman, 1910). He defined the result of this calculation, which he called a reliability coefficient, as the coefficient between one half and the other half of several measurements of the same thing (Spearman, 1910).

Subsequent developments of CTT involved proposals of various methods to perform this calculation (Cronbach, 1951; Guttman, 1945; Kuder & Richardson, 1947; Tryon, 1957). In general, the methods for calculating reliability coefficients in CTT assume that an observed score comprises one true and one error source of variance. To determine the amount of variance attributable to either source, CTT requires investigators to find parallel forms. In CTT, parallel forms are equivalent versions of the same tool for which differences in scores are attributable solely to what is called error in CTT. Methods for finding parallel forms include splitting a test in half or giving

different administrations of the same test at different times. To quantify the reliability of an instrument in CTT, investigators must calculate the correlation between the parallel forms. This correlation is known as a reliability coefficient (Brennan, 2000, 2010; Cronbach, 2004; Cronbach, Nageswari, & Goldine, 1963; Shavelson, Webb, & Rowley, 1989).

CTT is advantageous because the theoretical framework is conceptually simple and allows the calculation of a standard error of measure using the reliability coefficient (Brennan, 2010). However, error estimates and estimates of reliability vary according to the method used to calculate the reliability coefficient (Brennan, 2010; Shavelson et al., 1989). Because CTT conceptualizes only a single error term, it cannot distinguish between different sources of error (Brennan, 2010; Cronbach, 2004; Shavelson et al., 1989).

Generalizability theory. Generalizability theory (GT) employs a statistical method derived from Analysis of Variance (ANOVA) to allow investigators to distinguish between different sources of error variance that lead to inconsistencies in observed scores (Brennan, 2001, 2010; Cronbach et al., 1963; Shavelson et al., 1989). To accomplish this distinction, GT first requires investigators to specify the set of situations in which they will collect data and to which they will generalize scores. GT calls the set of situations from which researchers will collect data the universe of admissible observations and the set of situations to which they will generalize scores the universe of generalization.

Within the universe of admissible observations and the universe of generalization, GT allows measurement situations to be grouped into facets, which are similar sets of measurement conditions defined by researchers. Examples of sources of error variance commonly specified as facets are Occasions, Raters, and Items. In GT, investigators call the source of variance that corresponds to what is measured the object of measurement, and

investigators consider the object of measure true rather than error variance (Brennan, 2001, 2010; Cronbach et al., 1963; Shavelson et al., 1989).

GT distinguishes between two types of studies: generalizability (G) studies and dependability (D) studies (Brennan, 2001, 2010; Shavelson et al., 1989). Investigators complete G studies before D studies. In G studies, researchers estimate the variance contributed by specified facets and interactions to an individual observed score in the universe of admissible observations. In D studies, researchers estimate the variance of observed mean scores across facets and interactions in the universe of generalization. Researchers use variance estimates from G studies to estimate variance in D studies (Brennan, 2001, 2010; Shavelson et al., 1989).

Three additional considerations in G theory include whether a facet is random or fixed, whether facets are crossed or nested, and whether scores will be used for relative or absolute decisions. For the first consideration, a facet is fixed if the conditions covered by the facet in the study exhaust all possible conditions of interest. The term conditions in GT refers to the different levels of a facet (Brennan, 2001, 2010; Shavelson et al., 1989). For example, different conditions of a Settings facet might be different rooms within a school. If the Settings facet were fixed, then observed scores in a D study would be generalizable only to the rooms contained in the study. If the Settings facet were random, observed scores in a D study would be generalizable to other rooms. Whether a facet is fixed or random determines whether the variance attributed to the facet is considered error variance (Brennan, 2001, 2010; Shavelson et al., 1989). Fixing a facet decreases error variance and increases generalizability at the expense of a narrowed range of interpretations because conceptually fixing a facet restricts the universe of generalization (Brennan, 2001, 2010; Shavelson et al., 1989). Researchers can use variance estimates from G studies where facets are random in D studies where those same facets are

fixed, but researchers cannot use variance estimates from G studies with fixed facets in D studies where those facets are random (Brennan, 2001, 2010; Shavelson et al., 1989).

For the second consideration, facets are crossed if they are exposed to every condition of every other facet, and facets are nested if they are exposed to only some conditions of another facet. For example, in a design with raters and students, crossed facets occur when every rater rates every student, and nested facets occur when some raters rate some students but not others. Investigators can only use G studies with nested facets to estimate variance components for D studies where the same facets are nested. However, researchers can use variance estimates from G studies with crossed facets in D studies with nested facets, as long as the same facets are nested as in the G study (Brennan, 2001, 2010; Shavelson et al., 1989).

The last consideration involves whether the scores from the tool examined in the study will be used for relative or absolute decisions. Relative decisions involve the rank order of differences between individuals, and absolute decisions involve the comparison of a score to some sort of criterion. In relative decisions, investigators consider variance components associated with rank ordering individuals, such as interactions between facets and the object of measure, to be error. In absolute decisions, investigators consider all variance components, with the exception of the object of measure, to be error (Brennan, 2001, 2010; Shavelson et al., 1989).

Although variances components are a central focus of GT, GT also uses two types of coefficients similar to reliability coefficients in CTT (Brennan, 2001, 2010; Shavelson et al., 1989). These two coefficients are the generalizability coefficient and the index of dependability. Investigators use the generalizability coefficient to inform relative decisions, and they use the dependability coefficient to inform absolute decisions. As a result, investigators consider interactions between facets

and the object of measure sources of error in generalizability coefficients, and they consider all variance components, with the exception of the object of measurement, sources of error when calculating dependability coefficients (Brennan, 2001, 2010; Shavelson et al., 1989).

Overall, the central question GT tries to answer is the extent to which a score from a data collection tool is generalizable across a set of measurement situations defined by the researcher. By creating a framework through which researchers describe potential sources of error in their measurement situations, GT allows researchers to estimate more than one source of variance that might contribute to measurement error. (Brennan, 2001, 2010; Shavelson et al., 1989). For additional information on G theory and details on statistical procedures used to estimate variance components, please see Appendix C.

Validity

While CTT and classical conceptualizations of validity consider issues related to reliability and validity separately, GT and modern conceptualizations of validity overlap somewhat in the issues they cover (Brennan, 2000; Kane, 2013; Messick, 1995). Brennan (2000) distinguishes GT from the concept of validity by pointing out that although some facets commonly specified in GT also involve issues associated with reliability, such as score generalizations across raters and occasions, GT cannot answer questions regarding whether those facets are useful, meaningful, or true in any sense (Brennan, 2000). Such questions remain the concern of validity theories.

Validity refers to the degree to which evidence and theory supports an interpretation of a score derived from a data-collection tool (AERA, 2014). Broadly speaking, three major conceptions of validity exist: classical validity (APA, 1966), Messick's (1995) unitary conceptualization of validity, and the argument-based approach to validation (Kane, 2013). The argument-

based approach to validation is useful for highlighting issues addressed in the present study because it allows investigators to specify how unreliability in scores from a data-collection tool undermines the validity of inferences from those scores, such as interpretations used in classification decisions and decisions within multi-tiered systems of support.

Classical validity. Classical validity recognizes three different sources of evidence for validity: content-related validity evidence, criterion-related validity evidence, and construct-related validity evidence (APA, 1966; Merrell et al., 2012). Content-related validity evidence involves the extent to which a measure represents the domain about which conclusion are to be drawn. Criterion-related validity evidence involves the extent to which scores from an assessment tool correlate with other measures of the construct of interest in the expected directions. Construct-related validity evidence involves the degree to which the construct a data-collection tool attempts to measure explains performances on the tool (APA, 1966; Merrell et al., 2012).

Unitary concept and argument-based approach. Classical approaches to evaluating the validity of a data-collection tool consider validity separately from the reliability of a tool. In contrast, Messick (1995) and Kane (2013) offer two conceptualizations of validity that incorporate issues related to reliability into theoretical frameworks for evaluating validity. Messick conceptualizes validity as a unitary concept that involves six different aspects of construct validity. Issues related to reliability fall under the aspect of generalizability in Messick's conceptualization of validity. Kane conceptualizes validity as a system of arguments justifying interpretations of scores derived from a data-collection tool. In Kane's validity conceptualization, studies investigating the reliability of a data

collection tool support generalization inferences in validity arguments (Kane, 2013).

The six different aspects of construct validity in Messick's (1995) unitary conceptualization of validity are content, substantive, structural, external, consequential, and generalizability. Of most interest for the present study, the generalizability aspect involves concerns related to the reliability of a data-collection tool, such as the ability to generalize scores across raters and occasions (Messick, 1995). As a result, studies investigating the reliability of DBRs are needed to establish the validity of DBR score interpretations under Messick's unitary conceptualization of validity.

The argument-based approach to validation offers a somewhat simplified approach to validating a tool compared to Messick's (1995) unitary conceptualization of validity by focusing on the arguments used to justify score interpretations (Kane, 2013). In the argument-based approach to validity, researchers must construct two arguments to justify the validity of score interpretations: an interpretation use argument (IUA) and a validity argument (VA).

IUAs consist of a chain of inferences that connect scores derived from a data-collection tool to their proposed use (Kane, 2013). Types of inferences found in these arguments include scoring inferences, extrapolation inferences, theory-based inferences, decision inferences, and generalization inferences. Of most concern for the present study, generalization inferences involve issues related to the reliability of an assessment tool. These inferences are extrapolations from a sample of scores from a data-collection tool to other samples of scores from the tool taken under different conditions (Kane, 2013).

To establish the validity of interpretations from a data-collection tool, investigators also must construct a validity argument to evaluate the

plausibility of the chain of inferences composing the IUA. Validity arguments for using DBR ratings for decisions related to screening, progress monitoring, and classification must look at studies investigating the reliability of DBR ratings to evaluate whether evidence exists for generalizing scores across possible sources of error, such as raters. That is the primary goal of existing GT research with DBRs.

Generalizability and DBRs

GT is useful for investigating the reliability of DBRs because it allows researchers to measure more than one possible source of error (Brennan, 2001, 2010; Kane, 2013; Shavelson et al., 1989). Potential error sources of interest in assessments of behavior are: scorers or raters, items, time, setting, method, and dimension (Cone, 1977). Of those potential error sources, raters are of particular concern in DBR research (Christ et al., 2009). Rater variance is relevant to all proposed uses of DBRs because those uses involve both relative and absolute decisions based on DBR data, and rater variance undermines the reliability and validity of both relative and absolute decisions.

Rater variance affects relative decisions through interactions between raters and students being rated (e.g. difference between how raters rate individual students), and it affects absolute decisions through both variance attributable to raters (e.g. differences between how raters rate all students) and interactions between raters and students. Therefore, under the argument-based approach to validity, rater variance undermines inferences involving generalizations of DBRs across raters. Problems with such inferences may invalidate DBRs for screening purposes or for use as evidence in special education qualification decisions. In addition, rater variance also may restrict generalizations of scores from DBRs across rater contexts and invalidate inferences involving the effectiveness of an intervention in different contexts with different raters. In other words, high rater

variance would invalidate arguments generalizing a students' behavior in response to an intervention across different rater contexts, such as different classrooms with different teachers. Because of those potential threats to the validity of DBR score interpretations, researchers must investigate the extent to which DBRs are generalizable across raters to establish the defensibility of DBRs for roles such as screening, progress monitoring, and as evidence for high-stakes decisions (e.g., special education eligibility).

Generalizability of DBRs and Rater Variance

Generalizability studies of DBRs in both naturalistic and controlled environments across a variety of populations and behavioral constructs found large variances attributable to raters and the interactions between raters and students (Briesch et al., 2010; Chafouleas et al., 2007; Christ et al., 2010; Volpe & Briesch, 2012). For example, Chafouleas et al. (2007) found raters contributed 40% of the variance to scores of the construct Works to Resolve Conflicts and 20% of the variance to scores of the construct Interacts Cooperatively. In addition, Briesch et al. (2010) found the interaction between students and raters contributed 20% of the variance to scores of Academic Engagement (AE). That research undermines the validity of both relative and absolute decisions informed by DBR data; however, some scholars have suggested that training might be a useful tool to improve DBR reliability (Briesch et al. 2010; Chafouleas et al., 2007). In addition, because raters in Briesch et al. were required to rate 12 students and in Chafouleas et al. were required to rate 15 students, Briesch et al. suggested rater fatigue may have reduced rater agreement due to the number of students raters were required to rate in the studies.

In addition to research in natural school settings, other work evaluated the reliability of DBRs in controlled settings. Christ et al. (2010) investigated the generalizability of the DBR constructs Visual

Distraction and Active Manipulation using 6, 10, and 14 point DBR scales completed by 125 college students. The students rated video recordings of children completing a frustration task. For all scale lengths, raters contributed about 20% of the variance to the DBR scores. Volpe and Briesch (2012) investigated the generalizability of the behavioral constructs Academic Engagement (AE) and Disruptive Behavior (DB) using DBR single-item and multi-item scales completed by two male school psychology doctoral students. The doctoral students rated video recordings of eight elementary school students transitioning to and engaging in large-group instruction. The combination of raters and the interaction between students and raters contributed 9% of the variance to both DBR-SIS and DBR-MIS scores for AE. For DB, the combined variance of raters and the interaction between students and raters contributed 13% of the variance for DBR-SISs but only 4% of the variance for DBR-MISs. In addition, generalizability and dependability coefficients did not reach .80 for scores of DB from the DBR-SISs despite the addition of more rating occasions in D studies. Based on those results and the results of previous studies, Volpe and Briesch suggested DBR-SISs might not be dependable measures of behavior over realistic time frames for progress monitoring (Volpe & Briesch, 2012).

Because of the large sources of variance attributed to raters in the G studies of DBR, DBR researchers suggest that DBR data be interpreted only within raters rather than generalized across raters (Briesch et al., 2010; Chafouleas et al. 2007; Christ et al., 2010). Within-rater interpretations of DBR data might be useful for applications within school psychology practice such as problem solving in a single classroom with a single teacher. However, other proposed uses of DBR data, such as progress monitoring and screening, require generalizations across raters.

Training and DBR Accuracy

Due in part to the observed variance across raters, DBR researchers developed a web-based rater-training module (Chafouleas et al., 2014). To inform the development of the training module, DBR researchers conducted several studies to investigate the impact of DBR training on rater accuracy. The studies investigated several types of rater training across several DBR constructs and levels of student behavior.

Chafouleas et al., (2012) used video-recording to evaluate the impact of three types of training for two different lengths of time on the accuracy of DBR ratings of AE and DB among second and third graders engaged in three different base rates (e.g. levels) of each behavior. The training types were standard training (ST), frame-of-reference training (FOR), and frame-of-reference plus rater-error training (FOR + RET). In the ST group, participants read a book chapter describing DBRs, viewed a slideshow and lecture describing DBRs, and then practiced the DBR ratings and received feedback on their performance. In the FOR group, participants read the same book chapter, viewed the same didactic materials, and received feedback similar to the ST group, but unlike the ST group the feedback included specific concrete examples. Participants in the FOR + RET group viewed slides with examples of three types of response bias and rater errors in addition to the same practice and feedback as the FOR group. Chafouleas et al. also investigated the impact of training time by dividing participants into groups that received either three or six practice ratings.

Overall, the study results suggest a DBR training package increases the accuracy of DBR ratings for AE and DB. However, the effect of the number of training components appeared to depend on the base rates of the behavior. Additional training components did not increase rater accuracy for low base rates of either AE or DB. Further, additional training components decreased rater accuracy for medium rates of AE. For DB, rater accuracy appeared to

increase for medium and high base rates of the targeted behaviors following training. Finally, increased practice opportunities generally appeared to result in increased rater accuracy, especially for medium and high base rates of behavior (Chafouleas et al., 2012).

Building off Chafouleas et al. (2012), Chafouleas et al., (2014) examined the impact of a web-based training module with similar components as those used in the prior study. These components included (a) a didactic presentation familiarizing students with assessing behaviors using DBRs, (b) modeling that included frame-of-reference training, and (c) multiple opportunities to practice and receive immediate corrective feedback. The behavioral constructs rated were AE, RB, and DB. The results of the study suggest ratings from individuals who completed the training were more accurate than individuals who did not complete the training. However, as in previous studies, the impact of training on rater accuracy depended on the behavioral construct rated and the base rate of the behavior. Rater accuracy was significantly improved for high levels of AE and RB, but contrary to the researchers' hypothesis that the training would improve rating accuracy of medium levels of behavior, only ratings of medium levels of RB were more accurate (Chafouleas et al., 2014).

Other research provides limited support to the idea that training increases rater accuracy. Lebel et al. (2010) found raters who received both minimal and more extensive training produced accurate DBR ratings of AE and DB. However, raters who received the least amount of training produced the most accurate ratings of AE. For DB, there were no differences in rater accuracy between participants who received different training amounts. Harrison et al. (2014) also found participants who received three different training amounts produced accurate ratings of AE, DB, and RB. However, increased training improved rater accuracy for DB but not for AE and RB. The training impact also depended on behavior base rate. For AE, participants

rated medium and low behavior base rates significantly less accurately than high base rates. Contrary to expectations, the least trained participants rated low levels of AE more accurately than participants who received more training. For RB, participants rated medium base rates of the behavior significantly less accurately than low or high base rates of the behavior (Harrison et al., 2014).

Need for G Studies of DBRs by Trained Raters in Naturalistic Settings

Regardless of the outcome of DBR training research, it should be noted that existing work has only evaluated the impact of training on the accuracy of DBR ratings. Accuracy, as defined in that research, involves the relationship between DBR ratings and external criteria, such as SDOs and expert DBR ratings (Chafouleas et al., 2012; Chafouleas et al., 2014; Harrison et al., 2014; Lebel et al., 2010). In the classical conception of validity, the relationships explored in the existing studies on the impact of training on DBRs constitute evidence of criterion-related validity, which concerns the relationship between scores from a data collection tool and scores from other instruments measuring the same or different constructs (APA, 1966). In Messick's (1995) unitary conception of validity, the relationships in the accuracy studies would constitute evidence for the external aspect of construct validity, which concerns the relationship between performance on an assessment tool and performance on other assessment tools. In contrast, the reliability of DBRs concerns the generalizability of scores from data-collection tools in Messick's conception of validity. Likewise, in the argument-based approach to validation, studies examining the accuracy of DBR ratings provide backing for theory-based inferences derived from scores from a data-collection tool but not generalization inferences (Kane, 2013)

Further, the studies described in the previous subsection investigated the impact of training on rater accuracy in controlled laboratory conditions.

Differences between conditions encountered by raters in the laboratories and in schools might limit the generalizability of the accuracy studies to naturalistic settings in a school. For example, the length of the video clips used in the studies might limit generalizability of the studies because DBRs usually are completed over longer time periods when used in schools (Chafouleas et al., 2014). In addition, with the exception of Lebel et al. (2010), all of the studies used convenience samples of undergraduate students. Chafouleas et al. (2014) suggests such samples might not generalize to school-based settings because undergraduates have less incentive to master the training compared to professionals in a school. For all those reasons, studies examining the psychometric properties of DBRs completed by trained raters must be validated in the naturalistic setting of a school.

Summary and Present Study

Reliability is a critical piece of evidence regarding the defensibility of a data-collection tool as it concerns the consistencies and inconsistences of scores derived from the tool. In the context of DBRs, inconsistencies due to raters are of particular concern because of the proposed use of DBRs for instructional decision-making. Proposed uses for DBRs include (a) collecting data to design and evaluate interventions, (b) collecting data to screen students at risk for mental health problems, and (c) collecting data to monitor students' responses to interventions.

Inconsistencies due to raters limits the use of data gathered using DBRs to comparisons made within raters. Those comparisons are useful for problem-solving efforts in a single setting with a single rater; however, comparisons between raters are needed to use DBR data for screening and progress monitoring, as well as for high-stakes decisions associated with the traditional role of school psychology. For example, screening involves decisions regarding whether a student will receive additional supports. Low

agreement among teachers who use DBRs in their classrooms may cause students in one classroom to be more likely to receive these additional supports than students in another classroom because of differences in how the teachers rate their students rather than the students' behaviors. Regarding problem-solving, low DBR rater agreement may make it difficult to use data regarding a student's response to an intervention in one classroom to inform decisions regarding interventions a student may receive in another classroom. For example, score differences between the student's behaviors in each classroom might result from differences in how raters rate the student rather than differences in the student's behavior post intervention or differences in the fidelity of the implementation of the intervention. Such a difficulty may make the problem solving process less efficient and therefore require school psychologists and teachers to spend valuable additional time and resources problem solving a student's behavior. Moreover, such difficulties may delay the implementation of necessary and beneficial interventions.

Existing studies of DBRs using generalizability theory consistently find a large amount of variance due to raters. However, to date no studies investigated the generalizability of DBRs by raters who received a formal training module. Although existing research suggests training increases the accuracy of DBR ratings, the impact of training depends on the behavioral construct targeted, the training method used, and the criterion used to determine accuracy. Further, accuracy is a different construct than reliability, and the impact of DBR training on reliability is unknown.

The present study addresses the noted gaps in the literature by using GT to estimate the variance attributable to raters in two separate G studies conducted in the naturalistic setting of a school: a G study with raters exposed to training and a G study with raters not exposed to training. The present study compares the variances in these separate G studies to determine if the variance attributable to raters differs. In addition, because

previous research involved raters rating more than 10 students and as a result DBR researchers (i.e., Briesch et al., 2010) suggested rater fatigue may have reduced rater agreement, the present study replicates Briesch at al. with reduced rater demands by having raters rate less students.

Investigating the impact of training and rater fatigue on unwanted rater variance is an important step towards determining whether continuing efforts to develop methods to train raters in the use of DBRs are effective in reducing unwanted rater variance, as well as provide information to school psychologists interpreting DBRs regarding potential threats to the reliability and validity of DBR data. To summarize, the present study addresses the following questions:

1. How much variance is attributable to raters in the present G study with trained raters compared to previous G studies examining DBRs by untrained raters?
2. How much variance is attributable to raters in the G study with untrained raters compared to Briesch et al. (2010)?
3. How much variance is attributable to raters in the G study with trained raters compared to the G study with untrained raters?

In addition, the present study contains a focus group conducted at the end of the G studies. Focus groups can be useful for understanding how individuals arrive at conclusions (Stewart et al., 2016). Therefore, the focus group in the present study gathered information about how raters arrived at their DBR ratings. Such information may aid understanding of reasons behind rater disagreement, as well as suggest ways to reduce such disagreement. Focus groups also can be useful for gaining information about community needs (Liamputtong, 2011). As a result, the focus group in the present study gathered information about the views of teachers and paraprofessionals regarding the DBRs. That information suggests ways to improve the design of DBRs in terms of feasibility and utility, and the

information provides insight into potential problems with the respectful behavior construct. Moreover, information from the focus group also suggests potential improvements to current DBR items, as well as new items measuring additional constructs of potential use. Overall, the focus group accomplished two goals:

1. Inform the design of DBR items.
2. Inform the development of future constructs measured by DBRs.

Chapter 3: Method

Participants

The generalizability (G) studies took place in two kindergarten classrooms in a prekindergarten through eighth grade combined elementary and middle school located in a large urban district in the mid-Atlantic region of the United States. Participants in the studies included two kindergarten teachers and two paraprofessionals who served as raters in the G studies and nine students who were rated in the studies. Each classroom contained three separate G studies, one for each behavioral construct, and each G study had one teacher and one paraprofessional serving as raters.

Raters were recruited through in-person solicitation and fliers throughout the district. To be eligible for the study, raters needed to work in a pre-kindergarten, kindergarten, or first grade classroom in which two professionals agreed to participate in the study. Raters in the first classroom (Classroom 1) included one White female teacher with a master's degree and seven years teaching experience and one African American female paraprofessional with 30 years of experience. Raters in the second classroom (Classroom 2) included one White female teacher with a bachelor's degree and two years teaching experience and one African American male paraprofessional with three years of experience. Raters in the G studies also participated in the focus group component of the present study.

Although the teachers had experience completing behavior rating scales, none of the raters in the study had prior experience with DBRs or other direct behavior rating methods, such as SDOs. Raters were given 25-dollar gift cards for agreeing to participate in the study. For finishing the study and participating in the focus group, raters received additional 75-dollar gift cards.

Teachers in each classroom helped identify five students with a mixture of behavioral base rates (e.g. high or low levels of a given behavior). In

other words, students were selected for the G studies with the goal of creating a sample of students with a mixture or low, medium, and high levels of each behavior measured by the DBRS. Teacher impressions of students' behavior guided this process.

Of those students, parents of five students in Classroom 1 and parents of four students in Classroom 2 consented to allow their children to participate in the study. Student participants in Classroom 1 included one African American female, one White female, two Hispanic females, and one Hispanic male. Student participants in Classroom 2 included one White female, two Hispanic males, and one White male.

The Pennsylvania State University Institutional Review Board (IRB) approved procedures for gaining participants' assent and consent for the study, and written consent was obtained from teachers and paraprofessionals using forms and procedures approved by the IRB of Baltimore City Public Schools. Parents of student participants gave consent for their children's participation using forms and procedures approved by the Baltimore City Public School's IRB, and student participants gave assent using forms and procedures also approved by that IRB. Please see Appendix D for copies of the IRB approval letters from Penn State and Baltimore City Public Schools.

Materials

DBR forms. All raters completed Direct Behavior Rating Single Item Scale (DBR) forms containing items measuring three constructs: academic engagement (AE), disruptive behavior (DB), and respectful behavior (RB). Individual DBR items consisted of one 10-centimeter horizontal line divided into 10 equal segments by 11 much smaller vertical lines, underneath which appeared numbers ordered from 0 to 10 from left to right (see Appendix A). Faces depicting a frowning face, neutral expression, and smiling face were located below the numbers 0, 5, and 10 respectively on the items measuring AE and RB, and those faces appeared in the reverse order on the DB item. The

DBR forms contained definitions and examples of each behavioral construct on the top of the form along with overall directions instructing raters to mark the percentage of time the student engaged in each behavior (see Appendix A). Instructions on the forms also included a disclaimer that the percentages did not need to total 100 across each behavior because some behaviors could co-occur.

Focus group guide. Open ended questions appeared first in the interview guide followed by more specific questions per guidelines from Stewart et al., (2007). Also following those guidelines, question content aligned with research questions guiding the focus group (see Appendix B). For example, the interview used questions regarding rater agreement with behavioral construct definitions and the reasoning raters used to assign ratings to gather information about reasons for rater disagreement and potential problems with the RB construct suggested by research. Audio recording software on a personal computer equipped with a microphone recorded participants' conversations in the focus group along with audio recording software on a smartphone, which functioned as a backup recording device. The researcher transcribed recordings using word processing software on a personal computer.

Training module. Raters in Classroom 1 completed an online DBR training module prior to the study (University of Connecticut, 2014). The training module lasted about 45 minutes, and it is broken into three components. First, the module presented trainees with a narrated PowerPoint presentation that provided information about DBRs and the three behavioral constructs used in this study. Second, the module provided trainees with a video example of each behavioral construct and the correct rating for the example. Third, the module allowed trainees to rate additional video samples of students engaged in each behavior construct, after which it provided feedback regarding the trainees' ratings.

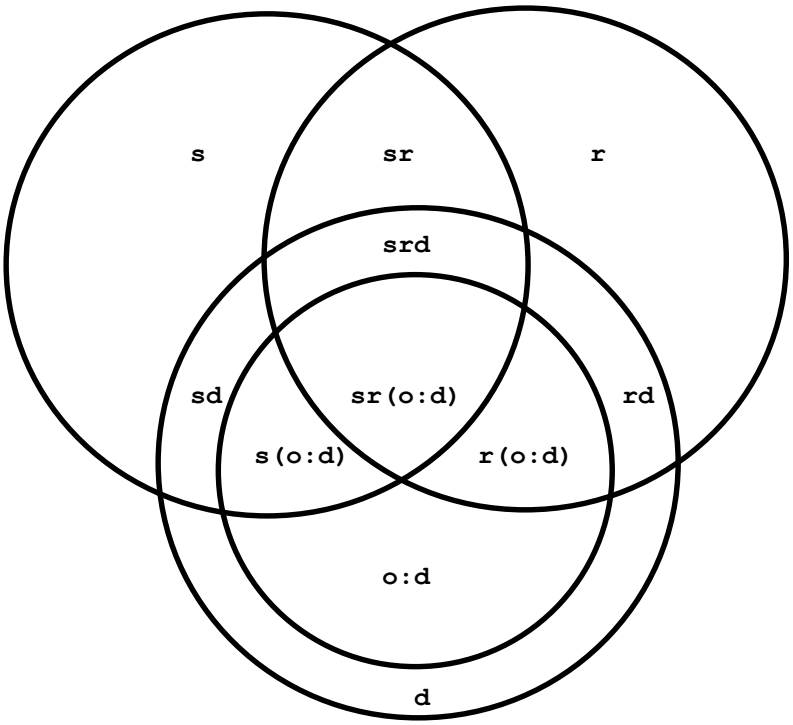
Study Procedures

The researcher met with all raters prior to the studies to highlight basic characteristics of DBRs. Raters in both Classroom 1 and Classroom 2 received enough information about DBRs to complete the DBR forms. The researcher summarized the directions on the DBR forms and the definition of each behavioral construct, pointed out the location of the directions and the behavior definitions on the DBR forms, and answered questions asked by the raters. Next, the researcher used a random number generator to randomly assign teachers and paraprofessionals from one of the classrooms to enroll in the additional training component, which consisted of the online module described above. The random number generator came from the website Random.org.

Generalizability (G) studies and focus group. Prior to the studies, the researcher worked with the teachers to identify three periods during the day that would be appropriate to complete DBR ratings. The researcher and raters decided ratings would be completed based on student behaviors occurring during the same 30 minute observation periods each day over the course of the 10 days of the study. The 30 minute observation periods constituted natural time periods during the day after which raters were able to take breaks to complete the DBR ratings. Two of those rating periods occurred in the morning, and one rating period occurred in the afternoon. The morning periods covered warm ups and phonics lessons, and the afternoon period covered math lessons. The researcher and the raters chose these time periods to match the activities during which DBR ratings occurred in Briesch et al. (2010) as closely as possible. The focus group occurred in Classroom 2 on the ninth day of the G studies.

Analysis Procedures

The designs of the G and D studies were Students crossed with Raters crossed with Occasions nested in Days (s X r X o:d). Because the G study design was unbalanced due to the presence of missing data, the researcher used UrGENOVA (Brennan, 2001) to estimate variance components for the G studies. UrGENOVA is a software program that uses the analogous ANOVA procedure (i.e. Henderson's Method 1) to compute variance components for G theory designs that are unbalanced due to nesting or missing data (Brennan, 2001). Please see Figure 1 below for a visual representation of the design of the G and D studies.



Variance Components

Object of Measure	Interactions
Students **(s)**	Students X Days **(sd)**
Fixed Facets	Raters X Days **(rd)**
Days (d)	Students X Occasions nested in Days **(s(o:d))**
Occasions nested in Days **(o:d)**	Raters X Occasions nested in Days **(r(o:d))**
Random Facets	Students X Raters X Days **(srd)**
Raters **(r)**	Students X Raters X Occasions nested in Days **(sr(o:d))**

Figure 1. Venn diagram of the variance components of the G and D studies. The D studies are mixed models (i.e., have fixed facets), and the G studies are random models (i.e., have all random facets). Each circle represents a facet or the object of measure, and the areas where the circles cross represent interactions. Circles completely enclosed by other circles represent nested facets.

The designs of the two G and D studies were Students crossed with Raters crossed with Occasions nested in Days. Students (s) were the object of measure, and Raters (r), Days (d), and Occasions nested in Days (o:d) were facets. The G studies were random models, meaning all facets were random. The D studies were mixed models, meaning some facets were fixed and others were random: Days and Occasions nested in Days were fixed, and Raters were random. One of the studies investigated the generalizability of DBRs

completed by trained raters, and the other study investigated the generalizability of DBRs completed by untrained raters. The variances attributed to raters and the interactions involving raters were compared across the two studies and with the Briesch et al. (2010) study.

Variance estimates from the G studies were used in D studies to measure the effect of additional rating days on the generalizability and dependability of the ratings. As recommended by Brennan (2001), GENOVA (Brennan & Crick, 1983) was used to perform the D studies. Consistent with recommendations from existing research, the facets of Occasions nested in Days and Days were fixed (Lei, Smith, & Suen, 2007). D studies yield two coefficients interpretable as estimates of measurement error: generalizability (G) coefficients and dependability (D or Phi) coefficients. The G coefficients measure relative error (i.e., differences in the rank order of students based on DBR ratings across each facet of the design). Phi coefficients measure absolute error (i.e., differences in the actual DBR ratings across each facet of the design). Values for G and Phi coefficients fall between 0 and 1, with values closer to 1 indicating less error compared to values closer to 0. Researchers commonly use .80 as a criterion for the minimum cut-off for acceptable use in roles such as screening and .70 for use in roles like progress monitoring (Salvia, Ysselydke, & Bolt, 2010). Because the facets of Days and Occasions nested in Days are fixed, variance associated with those facets does not contribute to error in the present D studies with the exception of interactions involving students or raters.

In the focus group, the researcher used thematic analysis (Liamputtong, 2011) to analyze transcriptions for recurring themes using methods from Liamputtong (2011) and Stewart et al., (2007). To categorize themes emerging from the focus group, the researcher developed a coding framework to organize participant responses. The coding framework consisted of the following

codes: feasibility, utility, training, constructs, electronics, and mechanics.

Chapter 4: Results

Descriptive Data

Across all behavioral constructs, ratings tended to fall at the extreme ends of the scales. Raters tended to report high levels of academically engaged and respectful behavior and low levels of disruptive behavior. As can be seen below in Table 1, ratings from the untrained raters in Classroom 2 fell closer to the extreme ends of the scales than ratings from the trained raters in Classroom 1.

Table 1

Descriptive Statistics of Direct Behavior Ratings

	Academic Engagement				Respectful Behavior				Disruptive Behavior			
	M	*SD*	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max
Trained	8.8	1.04	4	10	9.1	0.83	7	10	0.68	1.14	0	7
Untrained	9.8	0.73	5	10	9.7	1.06	5	10	0.41	1.31	0	7

Note. Direct Behavior Ratings from Classroom 1 and Classroom 2.

Because two raters rated five students three times per day over 10 days in Classroom 1, 300 total ratings were possible. Of those possible ratings, 10% of the values in the data matrix were missing. In Classroom 2, of the 240 possible ratings that could occur due to two raters rating four students three times per day over 10 days, 14% of the values in the data matrix were missing. These missing data amounts are comparable to prior DBR studies in naturalistic settings (Briesch et al. 2010; Chafouleas et al., 2007). Missing data occurred due to logistical difficulties inherent in collecting data in an educational environment. Those difficulties included student absences, teacher absences, and paraprofessionals being called to cover other classrooms for absent teachers. Missing data were not imputed but were instead handled using the analogous ANOVA procedure (i.e., Henderson's Method 1) in UrGenova (Brennan, 2001).

Generalizability Studies

Model descriptions. The G and D studies in each classroom estimated the amount of variance contributed to the DBR scores by each facet in the study: Students, Raters, Occasions nested in Days, and Days, along with the variance contributed by interactions between those facets. Interpretations for each variance component in the G and D study models are summarized in Table 2 on the next page.

Table 2

Interpretation of G and D Study Variance Components

Variance Component	Interpretation
Students (s)	variance contributed by differences between individual students' ratings across the other facets of the study
Raters (r)	variance contributed by differences between raters' ratings across the other facets of the study
Days (d)	variance contributed by differences between ratings taken on each day across the other facets of the study
Occasions nested in Days (o:d)	variance contributed by differences between ratings on different occasions within each day across other facets
Students X Days (sxd)	variance contributed by differences between individual students' ratings on different days of the study across the other facets
Students X Raters (sxr)	variance contributed by differences in raters' ratings of individual students across the other facets
Raters X Days (rxd)	variance contributed by differences between raters' ratings between days across the other facets
Student X Rater X Days (sxrxd)	variance contributed by differences between students' ratings by each rater on different days
Residuals	
Students X Occasions nested in Days (sxo:d)	variance unaccounted for by other effects in design confounded with error
Raters X Occasions nested in Days (rxo:d)	
Students X Raters X Occasions nested in Days(sxrxo:d)	

Variance contributed by the Students facet indicates DBR ratings differ between students. Because efforts were made to include students with different levels of each behavior, variance on this facet would be true variance because such variance would indicate the DBR ratings measured actual

differences in students' behaviors. As a result, students are considered the object of measure in the present studies and therefore do not contribute to error. In contrast, variance contributed by the Raters facet indicates differences between the two raters explain differences in DBR scores. Because two raters rating the same student's behaviors need to agree for DBRs to be a reliable data-collection tool, such variance contributes to error. Variance contributed by the Days facet indicates DBR scores differed between each day of the study. Because the Days facet is fixed in the D studies, such variance does not contribute to error. Last, variance contributed by the Occasions facet indicates DBR scores differed between the three different occasions within each day. However, because occasions are nested within days in the studies, variance associated with the Occasions facet indicates differences between the overall ratings of all students for each occasion within each individual day of the study. This is different from a fully crossed design in which differences in occasions are interpreted across all days. As with the Days facet, variance attributed to the Occasions nested in Days facet does not contribute to error because it is fixed in the present D studies.

The study designs also contain interactions between the facets. For instance, variance attributed to the interaction between Students and Raters indicates differences between the raters' ratings of individual students. Variance attributed to the interaction between Students and Days indicates differences between individual student's ratings on different days. Variance attributed to the interaction between Raters and Days indicates differences in raters' ratings of the behavior of all of the students on different days. Regarding the three-way interaction, variance attributed to the three-way interaction between Students, Raters, and Days indicates differences between raters' ratings of individual students on different days.

All interactions involving Students and Raters contribute to error in the present D studies. Such interactions constitute relative error, which involves error associated with generalizing the rank order of the students according to their DBR scores. In addition, variance associated with the interactions plus variance associated with the Raters facet constitutes absolute error, which involves all error associated with generalizing students' DBR scores.

Interactions involving the nested Occasions within Days facet are residual effects. These residual effects are confounded with the three-way interaction between Students, Raters, and Occasions nested in Days. That interaction also is confounded with all other sources of undifferentiated error in the design. All residual effects constitute undifferentiated sources of variance, or error, within the context of the G theory study designs (Brennan 2001). Large amounts of variance attributed to residual effects would provide evidence that the facets of the study design (e.g. Students, Raters, Occasions nested in Days, and Days) do not account for much variance in the DBR ratings.

Model results. Please see Table 3 for the results of the G studies along with a comparison of the present results to those observed by Briesch et al., (2010). The results of the G studies are presented as the percent of score variance contributed by each variance component of the design for each of the G studies.

Table 3

Percent Variance Contributed By Each Component of the Present Generalizability Studies and the Briesch et al. 2010 Generalizability Study

	Trained Raters (Classroom 1)			Untrained Raters (Classroom 2)			Briesch et al. 2010
Effect	AE	RB	DB	AE	RB	DB	AE
Students (s)	11	7	8	6	12	16	47
Days (d)	7	5	6	7	0	7	0
Occasions nested in Days (o:d)	5	7	0	6	1	1	2.5
Raters (r)	2	3	7	0	0	1	7.5
Students x Days (sd)	0	1	0	21	48	40	0
Students X Raters (sr)	6	10	5	4	6	1	20
Rater X Days (rd)	0	5	0	1	0	0	2
Student X Rater X Days (sdr)	11	0	19	0	4	0	3
Residual							
Student X Occasions:Days (so:d)	17	7	1	0	0	1	0
Rater X Occasions:Days (ro:d)	16	11	19	0	0	0	4
Student X Rater X Occasions:Days (sro:d)	24	44	35	55	29	34	13
P	0.76	0.55	0.66	0.69	0.84	0.95	.82
Φ	0.75	0.46	0.51	0.69	0.83	0.94	.77

Note: AE = Academic Engagement; RB = Respectful Behavior; DB = Disruptive Behavior.

For the trained raters in Classroom 1 rating AE, the Raters facet contributed only 2%, and the interaction between students and raters contributed only 6% of the variance in DBR ratings. The Students, Days, and the interaction between Students, Raters, and Days contributed larger shares

of variance; however, the residuals contributed the largest share of variance for AE DBRs in the present study.

For RB in Classroom 1, the Raters facet contributed only 3% of the DBR rating variance, but the interaction between Students and Raters contributed 10% of the variance. The somewhat large variance share attributed to the interaction between Students and Raters indicates that the two trained raters rated individual students' respectful behavior slightly differently. As with AE, the residuals accounted for a very large share of the variance for DBR ratings of RB from trained raters.

Regarding DB for trained raters in Classroom 1, the Raters facet accounted for a slightly larger share of the variance (7%) than it did for AE and RB, and the interaction between Students and Raters accounted for 5% of the variance. Those variance estimates indicate the trained raters rated students' disruptive behavior slightly differently. In addition, the interaction between Students, Raters, and Days accounted for 19% of the variance, the largest share of the variance apart from the residuals. The relatively large share of the variance associated with the three-way interaction indicates Raters rated individual student's disruptive behavior slightly differently on different days.

In Classroom 2 (untrained raters), the Raters facet did not contribute any DBR score variance for AE, and the interaction between Students and Raters contributed only 4%. As with the trained raters, the residuals contributed the largest share of the variance. Apart from the residuals, the interaction between Students and Days contributed the largest portion of the variance, suggesting DBR ratings of individual students differed between days across Raters and the nested Occasions. For RB, the Raters facet did not contribute any score variance, and the interaction between Students and Raters contributed 6%. As with AE, the interaction between Students and Days contributed the largest portion of the variance, followed by the Students

facet. Although the residuals contributed a large portion of the variance, unlike the other constructs, the residuals did not account for the largest amount of variance for RB. The Raters facet did not contribute any variance to DB in Classroom 2, and the interaction between Students and Raters only contributed 1%. The interaction between Students and Days accounted for the largest portion of the variance, followed by the residuals.

Dependability Studies

Table 4 summarizes the results of the D studies. For the D studies, G and Phi coefficients were estimated for observation periods consisting of 15, 20, 50, and 100 days in length, as well as for the initial 10 days of the study. Based on existing standards, .80 was used as an acceptability criterion for the G and Phi coefficients in the present studies for decisions related to screening, and .70 was used for decisions related to progress monitoring (Salvia, et al., 2010). The Days and Occasions nested in Days facets were fixed, so variance associated with those facets does not contribute to measurement error in the present studies with the exception of interactions involving students and raters; however, because those facets are fixed, DBR ratings from the present studies cannot be generalized beyond the days and occasions contained in the studies.

Table 4

D Study Results As Number of Days of Ratings Increased

		Numbers of Days Ratings Completed				
		10	15	20	50	100
Classroom 1 (Trained Raters)	Academic Engagement	P = 0.76 Φ = 0.75	P = 0.78 Φ = 0.77	P = 0.78 Φ = 0.77	P = 0.79 Φ = 0.79	P = 0.80 Φ = 0.80
	Respectful Behavior	P = 0.55 Φ = 0.46	P = 0.55 Φ = 0.46	P = 0.56 Φ = 0.47	P = 0.56 Φ = 0.48	P = 0.56 Φ = 0.48
	Disruptive Behavior	P = 0.66 Φ = 0.51	P = 0.70 Φ = 0.53	P = 0.71 Φ = 0.54	P = 0.74 Φ = 0.57	P = 0.76 Φ = 0.57
Classroom 2 (Untrained Raters)	Academic Engagement	P = 0.69 Φ = 0.69	P = 0.69 Φ = 0.69	P = 0.70 Φ = 0.70	P = 0.70 Φ = 0.70	P = 0.71 Φ = 0.71
	Respectful Behavior	P = 0.84 Φ = 0.83	P = 0.83 Φ = 0.82	P = 0.83 Φ = 0.82	P = 0.83 Φ = 0.82	P = 0.82 Φ = 0.82
	Disruptive Behavior	P = 0.95 Φ = 0.94	P = 0.96 Φ = 0.95	P = 0.96 Φ = 0.95	P = 0.97 Φ = 0.96	P = 0.97 Φ = 0.96

Note: Initial G study and D study contained 10 days of ratings. Facets Days and Occasions nested in Days are fixed.

For the trained raters in Classroom 1, the 10-day observation period from the G study resulted in Generalizability (G) and Dependability (Phi) coefficients of .76 and .75 respectively for AE. The G and Phi coefficients did not reach .80 until 100 days of observations. For RB, the G and Phi coefficients were .55 and .46 respectively after 10 days of observation. Both coefficients did not approach .80 after adding additional observation days—100 days of observation yielded G and Phi coefficients of only .56 and .46. Regarding DB, the G and Phi coefficients were .66 and .51 respectively after 10 days of observation. Neither coefficient approached .80 after 100 days observation, with a G coefficient of .76 and a Phi coefficient of .57 after that time.

In Classroom 2, with the untrained raters, AE yielded G and Phi coefficients of .69 after 10 days of ratings. Neither coefficient approached .80 after adding additional rating days, with both coefficients reaching only

.71 after 100 rating days. For RB, the G coefficient was .84 and the Phi coefficient was .83 after 10 observation days. However, the coefficients decreased with additional days of observation, with both coefficients reaching .82 after 100 days of observation. Regarding DB, the G and Phi coefficients were .95 and .94 respectively. After 100 observation days, these coefficients reached .97 and .96.

Focus Group

As discussed, prior to analyzing the focus group to find common themes among participants' responses, the researcher created a provisional coding framework based on the content of the interview guide and the research questions guiding the focus group. Following analysis, the coding framework was revised and participant responses were grouped into the categories depicted below in Figure 2.

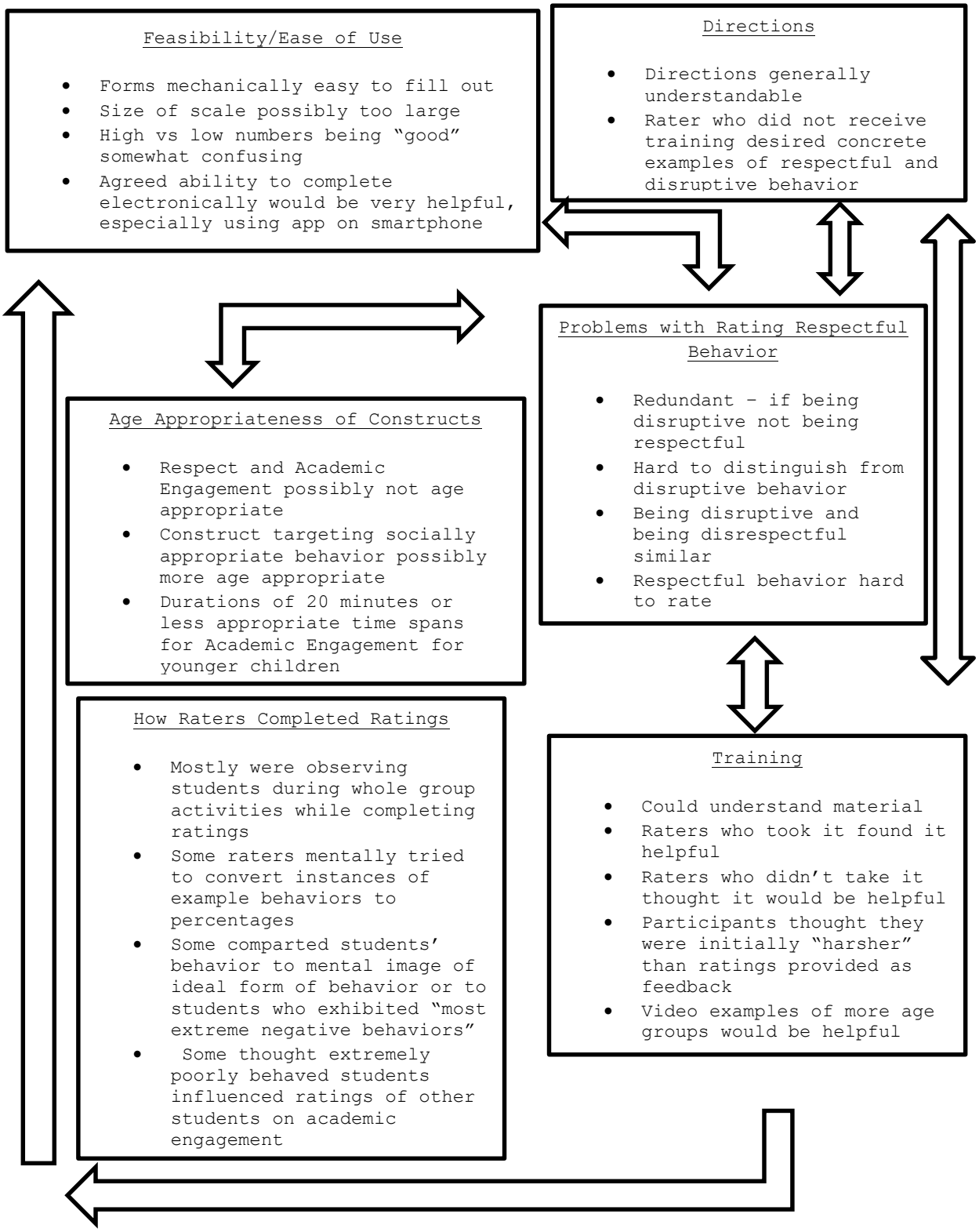


Figure 2. Thematic map of focus group conversation. Arrows indicate focus group participants brought up themes within connected boxes in connection with each other.

Overall, raters expressed similar impressions of the forms and process. When asked to give their general impressions of DBRs without additional guidance, raters reported that the forms were easy to complete mechanically; however, they expressed confusion with distinguishing between the RB and DB constructs. Some raters suggested that if students are being respectful they are not being disruptive. All raters felt the definitions of RB and DB were unclear, and all of the raters, even those who received the online training, expressed a desire for more examples of respectful and disruptive behavior. In addition, some raters felt that a 1 to 10 scale was too broad, expressing a preference for a 1 to 5 scale instead.

More focused questions regarding the definitions of the behavioral constructs revealed that the raters agreed with the definitions as they understood them. However, when asked to write examples of behaviors they believed fell under each construct, raters tended to provide different behaviors. For AE, behaviors listed by two or more raters include focus or being focused on something (e.g. work) and being on-task; one rater also listed working well with others as falling under AE. For RB, as with AE, two or more raters included following directions or classroom rules. Also in common with AE, one rater listed being on-task as an example of RB. Raters provided few common examples of DB, with calling-out or talking-out being the only example mentioned by at least two raters. Further, one rater differed from the other raters by emphasizing levels as distinguishing a behavior. For example, this rater listed high levels of being nice as an example of RB and high levels of being focused as an example of AE. Finally, some raters also suggested that a construct focused on social skills might be more appropriate than those focused on respect or academic engagement, especially for younger children.

When asked specifically to try to remember the reasoning behind their ratings, some raters reported they tried to correlate the number of instances

of each behavior they observed with a number for the percentage of time the student engaged in the behavior. Other raters stated they compared the behavior of the students they were rating to the behavior of other students in the room or to imaginary exemplars of a student exhibiting an ideal representation of a particular behavior. Some raters also mentioned that their estimate of the average behavior level of all children in their classroom influenced their ratings. The raters indicated they tended to switch between these strategies when completing their ratings, and none of the raters who received the online training mentioned a strategy derived from their training.

When reflecting on the training, raters who took the online module reported that it was helpful and understandable. Those who completed the training indicated that it was somewhat difficult to get their ratings to match the ratings provided by the module as feedback. Both trained raters felt that they were initially too harsh on the students in the examples based on the feedback they were given. One of the raters also suggested more examples be included from students belonging to additional age groups. Raters who did not take the training indicated that they thought training would have been helpful to them when completing their ratings. Finally, all of the raters said that electronically completing the ratings would be easier than completing them with pencil and paper. One rater volunteered that a smartphone app where a button with a smiley face could be pressed would be very convenient.

Chapter 5: Discussion

Training and DBR Rater Agreement

To investigate the impact of training on DBR rater agreement, the present G and D studies attempted to answer the three research questions posed in the introduction and at the end of the literature review. The first question asks: How much variance is associated with raters in the trained rater G study compared to the two previous G studies of untrained raters? Compared to one of the studies, Chafouleas et al. (2007), both trained and untrained raters in the two present G studies accounted for less variance in ratings of all three behavioral constructs, with the exception of trained raters rating disruptive behavior (DB). However, it should be noted that the behavioral constructs used in Chafouleas et al. (2007), Works to Resolve Conflicts and Interacts Cooperatively, were different than those used in the present G studies. As a result, a direct comparison with the present G studies is impossible.

The second question makes a more direct comparison between rater agreement in the present G studies and rater agreement in a past DBR G study by asking: How much variance is associated with raters in the current G studies compared to Briesch et al. (2010)? Briesch et al. (2010) is a more direct comparison to the present G studies because that study also used academic engagement (AE) as a behavioral construct, and that study also involved kindergarten students. Compared to Briesch et al. (2010), rater variances for both trained and untrained raters were lower in the present G studies. Also, although as noted direct comparisons are impossible, rater variances for both trained and untrained raters were lower for the other constructs rated in the present G studies than in Briesch et al. (2010), again with the exception of DB rated by trained raters.

Direct comparisons between DBR behavioral constructs are important because the present G studies attempt to investigate whether training impacts

DBR rater variance, and a secondary aim of the present studies is to investigate the impact of rater fatigue due to rater load (e.g., number of students rated) on rater agreement. Because rater variance differs between each behavioral construct, differences in rater variance caused by differences in the properties of the constructs themselves, such as the intelligibility of the construct definitions, might confound attempts to draw conclusions regarding the impact of training or fatigue. As a result, the third question driving the design of the present G studies asked: How much variance is associated with raters in the G study with trained raters compared to the G study with untrained raters?

Contrary to the expectation that training increases rater agreement for the trained raters compared to the untrained raters in the present studies, rater variance in the G study with untrained raters was lower than in the G study for trained raters. As a result, the present study does not provide evidence that training increases DBR rater agreement. However, because rater variance for both trained and untrained raters was lower than in previous G studies completed in schools with the exception of DB ratings by trained raters, the results of the present G studies provide evidence to support a suggestions from Briesch et al. (2010) that rater load may influence rater agreement.

In addition to the G and D studies, the present study also contained a focus group component in which raters were interviewed regarding their experiences using DBRs in the study. Rater responses in the focus group suggested possible reasons for differences between raters' DBR ratings in the studies. First, raters' descriptions of how they defined each DBR behavioral construct suggest differences in how raters defined each behavior. Second, raters, in particular the untrained raters, indicated they found it difficult to rate respectful behavior (RB). Related to that difficulty, the raters said they felt that if students were disruptive they were not being respectful,

suggesting difficulties rating RB and DB concurrently. Third, raters stated that they found the length of time they were required to rate students, 30 minutes, to be too long. Fourth, the raters suggested that a 10 point scale was too large and that they felt a 5 point scale would be easier to rate. Fifth, rater responses suggested they had trouble understanding the part of the instructions directing them to rate the "proportion of time" students engaged in each behavior.

Suggested Ways to Improve DBRS

The focus group also contained questions designed to elicit raters views regarding ways to make the instructions on the DBR forms more understandable, as well as ways to make DBRs more feasible to use in classrooms and more useful in the eyes of educators. Regarding the directions on the DBR forms, the raters suggested more examples of the behavioral constructs printed on the forms would be helpful. Regarding DBR feasibility, the raters suggested electronic versions of DBR forms, in particular on a smartphone, would be convenient for them to use. Regarding ways to make DBRs seem more useful to educators, the raters suggested a construct targeting socially appropriate behavior might be more useful than one targeting respectful behavior, in particular for younger children.

DBR Reliability

In addition to information provided by the present G studies and focus group that helps inform answers to the questions from the introduction and literature review, the present D studies provide generalizability estimates that help inform answers to questions related to the reliability of DBRs in different data collection roles.

Trained raters. Despite the increased rater agreement among the trained raters in the present G studies compared to Briesch et al. (2010), the G and Phi coefficients were lower in the present trained rater D study than in the Briesch et al. (2010) study. The larger share of variance contributed by the

residuals and the smaller variance share contributed by Students in the present study may have reduced the G and Phi coefficients compared to the Briesch et al. (2010) study. Although a direct comparison is impossible, of note, the G and D coefficients were markedly lower for RB and DB than for both AE in the present study and AE in Briesch et al. (2010). That difference suggests potential problems with those constructs, echoing prior research reporting potential problems with the RB construct in particular (Chafouleas et al., 2014; Riley-Tillman et al., 2009). Difficulties reported by raters in the focus group with rating RB and distinguishing between RB and DB may have reduced the overall reliability of the ratings.

Overall, because none of the coefficients approached .80 for trained raters except after 100 days of rating AE, the present study does not provide evidence that DBRs by trained raters are defensible for use in screening or for high stakes decisions. However, because both G and Phi coefficients were above .70 for AE, the present study suggests AE DBRs by trained raters may be defensible for use in progress monitoring and problem solving when that information is used for low-stakes decision-making.

Untrained raters. As with the trained raters, the G and Phi coefficients were lower for the untrained raters for AE than in the Briesch et al. (2010) study; however, the G and Phi coefficients also were much lower for AE than the G and Phi coefficients for the trained raters. The larger variance amount contributed by differences in how the students behaved on different days and the residual compared to the trained raters likely led to the lower coefficient values for AE. For RB and DB, the G and Phi coefficients were higher than all of the other coefficients for both the trained raters and for AE in the Briesch study. The larger amount of variance attributed to Students and the smaller amount attributed to the residuals may have led to the increased reliability observed in Briesch et al. (2010). As with the trained raters, because the G and Phi coefficients

were near .70, the present study suggests AE DBRs may be defensible for use in progress monitoring. However, although G and Phi coefficients for RB and DB DBRs were above .80, which is the acceptability cut-off for high-stakes decisions, potential confounds involving rater accuracy and base rate complicate interpretations of G and Phi coefficients from the DB and RB D studies.

Base Rate and Rater Accuracy

Two other possible explanations for increased rater agreement in the present G studies compared to past studies relate to the base rates of the behaviors measured by the DBR constructs in the sample of students chosen for the study and the accuracy of the ratings. Related to rater accuracy, the present study did not involve classical notions of validity. In other words, the design of the present study did not include a component, such as a comparison with SDOs completed by trained raters on the same students, to ensure the validity or accuracy of the DBR ratings. As a result, the possibility that raters in the present studies, in particular the untrained raters, guessed when completing their ratings cannot be ruled out.

Supporting the possibility that raters guessed, raters in the focus group, in particular the untrained raters, stated they felt some of the DBR behavior constructs were difficult to rate. Raters in the focus group, again in particular the untrained raters, singled out the RB construct as being difficult to rate and differentiate from DB, and all raters appeared to have difficulty with the instruction to rate "the proportion of time" students engaged in each behavior. Moreover, ratings in the present studies, especially those from the untrained raters, tended to fall at the extreme ends of the scales. In other words, raters in the present study tended to see students as being highly academically engaged and respectful and not very disruptive. This tendency might be evidence that raters, in particular the untrained raters, were guessing, or, the tendency might reflect the actual

behaviors of the students. Without an external validity check, either or both of these possibilities cannot be ruled out.

Studies that investigate reliability and validity concurrently are rare in the literature. However, of interest, the Briesch et al. (2010) study also contained a G study of SDOs completed by trained raters alongside the teachers who completed the DBR ratings. The SDOs in Briesch et al. might be thought of as an external check on the validity of the DBRs in that study. Similar to present study, large amounts of variance were attributed to residuals in the SDO G study; however, large variances attributed to the residuals were not observed in the Briesch et al. G study. As a result, the DBR ratings in Briesch et al. may not have been accurate compared to the SDO's in that study or the DBRs in the current studies. Further, and perhaps of greater importance, based on the large variances attributable to the residuals in the SDO G study, behavior ratings of groups of general education students might not vary in large amounts between Students, Days, and Occasions within those Days.

The possibility that behavior ratings of general education students do not vary to a great extent across the facets measured in the present study and in most past DBR G studies is important because of the potential impact of behavior base rate on score reliability estimates. Prior research suggest samples with low DBR score variability may result in higher estimates of generalizability and dependability compared to samples with higher variability (Briesch, Volpe, & Ferguson, 2013). In other words, if students engage in high base rates of academic engagement and respectful behavior or low base rates of disruptive behavior, raters are more likely to agree than if students engage in low base rates of academic engagement and respectful behavior and high base rates of disruptive behavior. As a result, reliability estimates based on samples of general education students without

documented social, emotional, and behavioral problems may not generalize to samples containing students with high levels of problem behaviors.

Impact of Rater Fatigue

As noted, because both the trained and untrained raters exhibited similar amounts of rater agreement, the present study does not furnish evidence for the contention that training increases rater agreement However, rater agreement was higher for both trained and untrained raters in the present studies than in past DBR G studies in naturalistic environments that required raters to rate more students. Chafouleas et al. (2007) required raters to rate 15 students, and Briesch et al. (2010) required raters to rate 12 students. Briesch et al (2010) suggested the rater demands in their study may have inflated rater disagreement because of the number of students raters rated in their study. In general, the results of the present study support that idea because raters in the present study rated less students and rater agreement was higher than in the past studies. In other words, the number of students that raters evaluate may affect agreement between raters. Increased rater loads seems to decrease rater agreement, and reduced rater loads seems to increase rater agreement. The influence of rater load on rater agreement may occur because raters rating too many students may become fatigued, which may affect the reliability and validity of their ratings. Future research is needed in this area, such as studies replicating the present G and D studies.

Other DBR Properties of Potential Interest Suggested By Present Study

In addition to issues related to DBR defensibility, feasibility, and utility, the present G and D study results also suggest several other DBR properties of potential interest. First, when rater disagreements did occur, raters' scores tended to fall within one to two points from each other on the DBRs. To help illustrate the similarity in ratings from different raters, exact scores were pulled from the data file, and those ratings are displayed below in Table 5 and Table 6.

Table 5

Example Ratings from Untrained Raters in Classroom 2

Academic Engagement		Respectful Behavior		Disruptive Behavior	
Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2
10	10	10	10	0	0
8	8	5	5	7	5
10	10	10	10	0	0
10	10	10	10	0	2

Table 6

Example Ratings from Trained Raters in Classroom 1

Academic Engagement		Respectful Behavior		Disruptive Behavior	
Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2
10	10	10	9	0	0
9	9	10	9	1	0
9	9	9	9	1	0
9	8	10	9	1	1
10	9	10	9	0	1

Second, as seen in Table 1, the untrained tended to rate students more positively than the trained raters. In other words, the untrained raters rated the students in their classroom as being more academically engaged, more respectful, and less disruptive than the trained raters rated the students in their classroom. Training might make raters more confident or more inclined to look for ways to differentiate their ratings: in other words, to look for instances when students are academically disengaged, not respectful, or disruptive. However, as discussed, despite that difference, ratings from both trained and untrained raters tended to fall at extreme ends of the DBR scales. In other words, all raters tended to rate students as being very academically engaged, respectful, and not very disruptive. As noted, this property of the ratings may indicate problems with the validity

of the ratings or it may be a true reflection of the behavior of students in the samples.

Implications for Practice

The results of the present study have several implications for the use of DBRs in school psychology practice, as well as for future efforts to improve DBRs. First, the results of the present D studies suggest DBRs of AE may be defensible in low stakes decisions because G and Phi coefficients were above or close to .70 for both trained and untrained raters. Second, the G studies provide some evidence that decreased rater load may reduce error due to interrater disagreement. As a result, school psychologists may need to consider rater load when interpreting DBR data and working with teachers. Third, the results of the present G and D studies do not provide evidence that the online DBR training module alone reduces rater disagreement and therefore error. Until evidence emerges to support training reducing rater disagreement, school psychologists should exercise caution when drawing inferences from DBR data that require generalizations across raters, such as those related to screening or high-stakes decisions. As with any data-collection tool, data from DBRs should be evaluated alongside data from other sources when making decisions.

Psychologists in practice might also consider ongoing coaching strategies derived from efforts to improve intervention fidelity in behavioral consultation, such as performance feedback training, to augment online training for DBRs (e.g., Noell et al., 2011). Such strategies may be more effective for reducing error due to rater disagreement than individual training alone. Finally, participants' focus group responses suggest difficulty with rating respectful behavior, especially in the absence of training and when rated alongside disruptive behavior. Therefore, school psychologists should both ensure raters have training before rating respectful behavior and also be sure to answer any questions raters have

about the RB construct. Extra caution when interpreting ratings of RB also may be warranted.

Limitations

For theoretical reasons, DBR generalizability is limited to the time periods in which the ratings occur. DBRs share this limitation with other methods of direct behavior measurement (Lei et al., 2007). In the face of this limitation, researchers can only investigate the reliability of such instruments through repeated studies in different measurement conditions along with replications of those studies.

Comparisons of variance estimates and coefficients across direct behavior measurement G studies may be complicated by the following known difficulties. First, variance estimates and coefficients differ across behavior constructs. As a result, direct comparisons between the present study and past studies were limited to Briesch et al. (2010), which also used AE as a behavioral construct. However, despite diligent attempts to replicate the design and procedures of Briesch et al. as closely as possible, the demographics of the student and rater samples in the present studies differed, and logistical difficulties necessitated slight changes to procedures for collecting the ratings. Regarding demographics, the present studies took place in large urban district rather than a suburban town. In addition, the sample of raters in the present studies consisted of both teachers and paraprofessionals, rather than just teachers.

Regarding procedures for completing ratings, the most notable changes were lengthened rating periods for each occasion in the present studies and a larger spread of occasions throughout the day. In other words, although the ratings in the present studies took place at the same time each day, unlike Briesch et al., the occasions were spread throughout the day as opposed to occurring immediately after one another. The rating periods were structured that way so the raters could complete their ratings over occasions covering

the same or similar activities as those that occurred during the rating periods of Briesch et al. (2010).

In addition, as noted, base rates of behavior in a sample are known to affect measures of generalizability and dependability (Briesch, et al., 2013). As a result, the low score variability in Classroom 2 may have led to an overestimate of the generalizability and dependability of RB and DB. Also related to the G study in Classroom 2, raters may have produced inaccurate ratings due to troubles using and distinguishing between the constructs of RB and DB. As discussed, these accuracy problems may have become confounded with measures of reliability for those constructs. Because the present study did not include another rating method as a validity check in terms of classical notions of validity, such a possibility cannot be ruled out.

Because of these limitations, researchers must continue to investigate the reliability of DBRs and replicate existing studies of DBR reliability. These studies need to be completed for each DBR construct. In addition, an external criterion to check the accuracy of the DBR ratings would be useful to detect potential rater inaccuracies that might complicate generalizability and dependability estimates.

Future Directions

The present study suggests future directions for reliability research of DBRs, as well as potential future avenues for improving DBRs in terms of reliability, as well as utility, feasibility, and acceptability. For future reliability research, studies investigating DBR generalizability and dependability with similar rater loads are needed to replicate the findings from the current studies. Likewise, additional research is needed to evaluate the generalizability and dependability of DBRs by trained raters when those raters are responsible for rating a larger number of students. In addition, because low variance in students' behaviors may have confounded generalizability and dependability estimates in the present studies, future G

studies with small samples containing both students with externalizing behavior problems identified via other assessment methods and general education students would useful. Finally, future studies might investigate the effect of training followed by strategies, such as performance feedback training, for reinforcing and maintaining behavior changes induced by the training. Training modules along with ongoing coaching may be more effective than individual training alone.

Regarding ways to improve DBRs, participants' focus group responses suggest several possible avenues. First, additional examples of behaviors falling under each construct may be added to the DBR forms to grant requests from the focus group participants for additional behavior examples. These examples might constitute an exhaustive list of behaviors falling under each construct, which could improve rater agreement. Second, because raters spontaneously suggested a construct targeting socially appropriate behavior, such a construct may be more intuitive for educational professionals to rate. Third, the raters also indicated they used several different cognitive strategies to estimate the proportion of time students were engaged in each behavior when completing their ratings. Future research might investigate cognitive strategies likely to result in accurate ratings, and instruction of those strategies might be included in future DBR training modules. Such training might encourage raters to use similar cognitive strategies for rating students, thus increasing rater agreement.

Finally, because focus group participants enthusiastically endorsed the idea of completing DBRs electronically, researchers might create computerized versions of DBR forms, such as versions completed through a smartphone app, to increase the feasibility of DBRs for use in classrooms. Such electronic DBR forms might possess several advantages. Computerized DBR forms may reduce rater load and increase the feasibility of DBR-MIS, which may possess superior psychometric properties to DBR-SISs (Volpe & Briesch, 2012).

Computerized DBR forms may also have the potential to reduce the workload involved with entering DBR ratings for purposes of analysis. In addition to potentially enabling the easy creation of visual displays of DBR data, which would be helpful for problem solving, computerized DBR forms also may increase the feasibility of performing generalizability studies. Software for completing generalizability and dependability studies may be incorporated into software for collecting DBR ratings. That software might reduce the difficulties inherent in investigating the reliability of direct measurements of behavior by reducing differences between the contexts of the studies and the contexts in which the ratings occur (Lei et al., 2007).

Conclusion

To summarize, the results of the present G studies provide some support for the notion that the characteristics of the sample to be rated (e.g., number of students) may matter more than the impact of additional training. In addition, the results of the present D studies suggest AE DBRs may be defensible for use in low-stakes decisions, such as those related to problem-solving and progress monitoring. Last, rater feedback from the focus group suggests several potential reasons for rater disagreement when using DBRs and that teachers may consider other constructs (e.g., social skills) to be easier to rate and of more interest than the RB construct. Finally, efforts to modify and evaluate DBRs may benefit from modifications to DBR instructions, DBR training programs, and the mechanism used to record ratings. As discussed, those modifications might include more detailed instructions, following up training, instruction in cognitive methods used to complete ratings, and the use of electronic devices to facilitate the use of DBRs in practice and for research purposes.

References

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

American Psychological Association (2010). *Ethical principles of psychologists and code of conduct.* Retrieved from American Psychological Association website http://www.apa.org/ethics/code/principles.pdf

American Psychological Association Presidential Task Force on Evidence-Based Practice (2006). Evidence-based practice in psychology. *American Psychologist, 61*, 271-285. doi: 10.1037/0003-066X.61.4.271

American Psychological Association, American Education Research Association, & National Council on Measurement in Education (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

Brennan, R. L. (2000). (Mis) conceptions about generalizability theory. *Educational Measurement Issues and Practices*, *19*, 5-10. doi: 10.1111/j.1745-3992.2000.tb00017.x

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Spring-Verlag New York, Inc.

Brennan, R. L. (2001). Manual for UrGENOVA v. 2.1. Iowa City, IA: Iowa Testing Program.

Brennan, R. L. (2001). UrGENOVA [Computer Software]. Iowa City, IA: Iowa Testing Program.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*, 1-21. doi: 10.1080/08957347.2011.532417

Brennan, R. L. & Crick, J. E. (1983). GENOVA [Computer Software]. Iowa City, IA: The American College Testing Program.

Briesch, A. M., Chafouleas, S. M., & Riley-Tillman, C. T. (2010). Generalizability and dependability of behavior assessment methods to estimate academic engagement: A comparison of systemic direct observation and direct behavior rating. *School Psychology Review*, *39*, 408-421.

Briesch, A. M., Volpe, R. J., & Ferguson, T. D. (2014). The influence of student characteristics on the dependability of behavioral observation data. *School Psychology Quarterly*, *29*, 171-181. doi: 10.1037/spq/0000042

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*, 296–322. doi: 10.1111/j.2044-8295.1910.tb00207.x

Chafouleas, S. M. (2011). Direct behavior rating: A review of the issues and research in its development. *Education and Treatment of Children*, *34*, 575-591. doi: 10.1353/etc.2011.0034

Chafouleas, S. M., Christ, T. J., Riley-Tillman, T. C., Briesch, A. M., & Chanese, J. A. M. (2007). Generalizability and dependability of direct behavior ratings to assess social behavior or preschoolers. *School Psychology Review*, *36*, 63-79.

Chafouleas, S. M., Jaffery, R., Riley-Tillman, C. R., Christ, T. J., & Sen, R. (2013). The impact of target, wording, and duration on rating accuracy for direct behavior rating. *Assessment for Effective Intervention*, *39*, 39-53. doi: 10.1177/1534508413489335

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. M., Kilgus, S. P., Riley-Tillman, C. T., Jaffery, R., & Harrison, S. (2012). Preliminary evaluation of various training components on accuracy of direct behavior ratings. *Journal of School Psychology*, *50*, 317-334. doi: 10.1016/j.jsp.2011.11.007

Chafouleas, S. M., Riley-Tillman, T., Jaffery, R., Miller, F. G., & Harrison, S. E. (2014). Preliminary investigation of the impact of a web-based module on direct behavior rating accuracy. *School Mental Health*, 7, 92-104. doi: 10.1007/s12310-014-9130-z

Chafouleas, S. M., Volpe, R. J., Gresham, F. M., & Cook, C. R. (2010). School based behavioral assessment within problem solving models: Current status and future directions. *School Psychology Review*, *39*, 343-349.

Christ, T. J., Riley-Tillman, T. C., & Chafouleas, S. M. (2009). Foundation for the development of direct behavior rating (DBR) to assess and evaluate student behavior. *Assessment for Effective Intervention*, *34*, 201-213. doi: 10.1177/1534508409340390

Christ, T. J., Riley-Tillman, T. C., Chafouleas, S. M., & Boice, C. H. (2010). Direct behavior ratings (DBR): Generalizability and dependability across raters and observations. *Educational and Psychological Measurement*, *70*, 825-843. doi: 10.1177/0013164410366695

Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy*, *8*, 411-426. doi: 10.1016/S0005-7894(77)80077-4

Crick, J. E., & Brennan, R. L. (1983). Manual for GENOVA: A generalized analysis of variance system. Iowa City, Iowa: The American College Testing Program.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of the tests. *Psychometrika, 16,* 297-333. doi: 10.1007/BF02310555

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 116-127. doi: 10.1037/h0076829

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64,* 391-418. doi: 10.1177/0013164404266386

Cronbach, L. J., Nageswari, R., & Goldine, C. G. (1963). Theory of generalizability: A liberalization of reliability theory. *The British Journal of Statistical Psychology, 16,* 137-163. doi:10.1111/j.2044-8317.1963.tb00206.x

Deno, S. L. (2002). Problem-solving as "best practice." In A. Thomas & J. Grimes (Eds.). *Best practices in school psychology IV* (Vol. 1 pp. 37-55). Bethesda, MD: National Association of School Psychologists.

Evans, S. W., & Owens, J. S. (2010). Behavioral assessment within problem-solving models: Finding relevance and expanding feasibility. *School Psychology Review, 39*, 427-430.

Federal Interagency Forum on Child and Family Statistics. (2013). *America's children: Key national indicators of well-being.* Washington, DC: U.S. Government Printing Office. Retrieved from: http://www.childstats.gov/pdf/ac2013/ac_13.pdf

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255-283. doi: 10.1007/BF02288892

Harrison, S. E., Riley-Tillman, C. T., & Chafouleas, S. M. (2014). Direct behavior ratings: Considerations for rater accuracy. *Canadian Journal of School Psychology, 29,* 3-20. doi: 10.1177/0829573513515424

Herman, K. C., Riley-Tillman, C. T., & Reinke, W. M. (2012). The role of assessment in a prevention science framework. *School Psychology Review, 41,* 306-314.

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review, 34,* 507-519.

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, 4,* 435-451. doi: 10.1023/A:1025045412689

Hoagwood, K., & Johnson, J. (2003). School psychology: A public health framework I: From evidence-based practices to evidence-based policies. *Journal of School Psychology, 41,* 3-21. doi: 10.1016/S0022-4405(02)00141-3

Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C. § 1400 (2004).

Jacob, S., Decker, D. M., & Hartshorne, T. S. (2011). *Ethics and law for school psychologists* (6th ed.). Hoboken, NJ: John Wiley & Sons Inc.

Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (2016). *Handbook of response to intervention: The science and practice of multi-tiered systems of support* (2nd ed.). Boston, Massachusetts: Springer US. doi:10.1007/978-1-4899-7568-3

Kane, (2013). The argument-based approach to validation. *School Psychology Review, 42,* 448-457.

Kratochwill, T. R. (2007). Preparing psychologists for evidence-based practice: Lessons learned and challenges ahead. *American Psychologist, 62,* 829-843. doi: 10.1037/0003-066X.62.8.829

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151-160. doi:10.1007/BF02288391

Lebel, T., 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 Behavior Interventions, 12,* 55-63. doi: 10.1177/1098300708325265

Lei, P.W., Smith, M., and Suen, H. K. (2007). The use of generalizability theory to estimate data reliability in single-subject observational

research. *Psychology in the Schools, 44,* 433-439. doi: 10.1002/pits.20235

Lewis, T. J., Barbara, S., Mitchell, D., Bruntmeyer, T., & Sugai, G. (2016). School-wide positive behavior support and response to intervention: system similarities, distinctions, and research to date at the universal level of support. In S. Jimerson, M. Burns, A. VanDerHeyden (Eds.). *Handbook of response to intervention: The science and practice of multi-tiered systems of support* (pp. 703-717) (2nd ed). doi: 10.1007/978-1-4899-7568-3

Liamputtong, P. (2011). *Focus group methodology: Principals and practice.* London, UK: Sage Publications Inc.

Merrell, K. W. (2002). Social-emotional intervention in schools: Current status, progress, and promise. *School Psychology Review, 31*, 143-147.

Merrell, K. W. (2011). *Behavior, social, and emotional assessment of children and adolescents* (3rd ed.). New York, NY: Routledge Taylor & Francis Group.

Merrell, K. W. & Ervin, R. A., & Peacock, G. G. (2012). *School psychology for the 21st century* (2nd ed.). New York, NY: The Guilford Press.

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

New Freedom Commission on Mental Health. (2003). *Achieving the promise: Transforming mental health care in America. Final Report* (DHHS Pub No. SMA-03-3832). Rockville, MD: 2003. Retrieved from: http://govinfo.library.unt.edu/mentalhealthcommission/reports/FinalReport/downloads/FinalReport.pdf

National Association of School Psychologists (2010). *Principles for professional ethics.* Retrieved from

https://www.nasponline.org/standards-and-certification/professional-ethics

Noell, G. H., Witt, J. C., Slider, N. J., Connell, J.E., Gatti, S. L. Williams, K. L., ... Duhon, G. J. (2005). Treatment Implementation Following Behavioral Consultation in Schools: A comparison of three follow-up strategies. *School Psychology Review, 34,* 87-106.

Haahr, M., & Haahr, S. (2016). Random.org. http://www.random.org

Rehabilitation Act of 1973, 29 U.S.C. § 794 (1973).

Riley-Tillman, C. 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. doi: 10.1037/a0015248\

Salvia, J., Ysselydke, J. E., & Bolt, S. (2010). *Assessment in special and inclusive education (*11th ed.*)*. Boston, MA: Houghton Mifflin

Shavelson, R. J., Webb, N. M., Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44,* 922-932. doi: 10.1037/10109-000

Stewart, Shamdasani, & Rook (2007). *Focus groups.* Thousand Oaks, CA: Sage Publications Inc.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72-101. doi: 10.2307/1412159

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271-295. doi: 10.1111/j.2044-8295.1910.tb00206.x

Strein, W., Hoagwood, K., & Cohn, A. (2003). School psychology a public health perspective I: Prevention, populations, and systems change. *Journal of School Psychology, 41,* 23-48. doi: 10.1016/S0022-4405(02)00142-5

Stoiber, K. C., & Maribeth, G. (2016). Multi-tiered systems of support and evidence-based practices. In S. Jimerson, M. Burns, A. VanDerHeyden,

(Eds.). *Handbook of response to intervention: The science and practice of multi-tiered systems of* support (2nd ed.) [Adobe Digital Editions version]. doi: 10.1007/978-1-4899-7568-3

Sugai, G., & Horner, R. H. (2009). Responsiveness to intervention and positive behavior supports: Integration of multi-tiered systems approaches. *Exceptionality, 17,* 223-237. doi: 1080/09362830903235375

The National Advisory Mental Health Council Workgroup on Child and Adolescent Mental Health Intervention Development and Deployment. (2001). *Blueprint for change: Research on child and adolescent mental health.* Washington, D.C.: 2001 Retrieved from National Institute of Mental Health website: http://wwwapps.nimh.nih.gov/ecb/archives/nimhblueprint.pdf

The National Association of State Mental Health Program Directors, The Policymaker Partnership for Implementing IDEA, and The National Association of State Directors of Special Education. *(2000). Mental health, schools and families working together for all children and youth: Toward a shared agenda: A concept paper.* Retrieved from the IDEA Partnership website: http://www.ideapartnership.org/documents/Shared%20Agenda_final.pdf

Tilly, D. W., III. (2008). The evolution of school to science-based practice: Problem-solving and the multi-tiered model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V (pp. 17-36).* Bethesda, MD: National Association of School Psychologists.

Tryon, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin, 54,* 229-249. doi: 10.1037/h0047980

University of Connecticut (2014). *Direct behavior ratings: Training site.* http://www.directbehaviorratings.com/training/

U.S. Department of Health and Human Services (1999). *Mental health: A report of the Surgeon General.* Rockville, MD: 1999. Retrieved from United States National Library of Medicine website: https://profiles.nlm.nih.gov/ps/retrieve/ResourceMetadata/NNBBHS

U.S. Public Health Service. (2000). *Report of the surgeon general's conference on children's mental health: A national action agenda* (ISBN No. 0-16-050637-9). Washington, D.C.:2000. Retrieved from Surgeon General website: https://www.ncbi.nlm.nih.gov/books/NBK44233/

Volpe, R. J. & Briesch, A. M. (2012). Generalizability and dependability of single-item and multiple-item direct behavior rating scales for engage and disruptive behavior. *School Psychology Review, 41,* 246-261.

Appendix A

DBR Form

Direct Behavior Rating (DBR) Form: 3 Standard Behaviors

Date: M T W Th F	Student: / Rater:	Activity Description:
Observation Time: Start:_____ End: _____ ☐ Check if no observation today	Behavior Descriptions: **Academically engaged** is actively or passively participating in the classroom activity. For example: writing, raising hand, answering a question, talking about a lesson, listening to the teacher, reading silently, or looking at instructional materials. **Respectful** is defined as compliant and polite behavior in response to adult direction and/or interactions with peers and adults. For example: follows teacher direction, pro-social interaction with peers, positive response to adult request, verbal or physical disruption without a negative tone/connotation. **Disruptive** is student action that interrupts regular school or classroom activity. For example: out of seat, fidgeting, playing with objects, acting aggressively, talking/yelling about things that are unrelated to classroom instruction.	

Directions: Place a mark along the line that best reflects the percentage of total time the student exhibited each target behavior. Note that the percentages do not need to total 100% across behaviors since some behaviors may co-occur.

Academically Engaged

% of Total Time

0 1 2 3 4 5 6 7 8 9 10
0% 50% 100%
Never Sometimes Always

Respectful

% of Total Time

0 1 2 3 4 5 6 7 8 9 10
0% 50% 100%
Never Sometimes Always

Disruptive *

% of Total Time

0 1 2 3 4 5 6 7 8 9 10
0% 50% 100%
Never Sometimes Always

* Remember that a lower score for "Disruptive" is more desirable.

Appendix B

Focus Group Interview Guide and Provisional Coding Framework

DBR Follow Up Study Focus Group Interview Guide

Interviewer: First of all, I want to thank you all again for participating in this study. The information from your ratings will help to inform the development of future rating tools. For this interview, there are no right or wrong answers to the questions; rather, this interview is meant to gather information regarding how professionals using DBRs view DBRs and how they complete their ratings. As a result, it is important that you all give your honest feelings regarding how you all view DBRs and how you completed the ratings. Please feel free to give your answers to each of the questions after they are asked and to build off the answers of other participants if those answers cause you to think of something to add.

First, please describe your overall feelings regarding DBRs.

Prompt(s): Did you find them easy to use?

Were you able to complete them easily during the course of the school-day?

Second, please describe your feelings regarding the behavioral constructs you rated.

Prompt(s): Do you think those are important sets of behaviors to monitor?

Were the constructs easy to understand?

Do you agree with the definitions of the behavior constructs?

Do you agree with the examples of specific behaviors for each construct? Do you agree that those behaviors fall under the construct?

Third, please describe your feelings regarding the instructions on the DBR forms.

Prompt(s): Were the overall instructions easy to understand? Did you understand what was meant by rating "the percentage of time the student exhibited the target behavior"?

Prompt: Did you find the DBR items themselves easy to use? Did you find the definitions of the behavioral constructs easy to understand? Were the examples of specific behaviors falling under the groups of behaviors for (say construct) construct helpful?

Prompt: Do you agree with those examples?

Fourth, please describe some ways that you think the DBRs can be improved?

Prompt(s): How would improve the constructs?

How would you improve the instructions on the forms?

How would you improve the definitions of the behavioral constructs?

What are some ways that the DBRs could be improved to make them easier for you to use?

Would completing the ratings electronically (e.g. on a computer, tablet, smartphone help)?

Now I am going to ask you to write specific examples of behaviors you thought fell under each construct on these pieces of paper.

(passes out papers)

First, for academically engaged behavior, please tell me some of the specific behaviors that you considered indicative of academic engagement when completing your ratings.

For respectful behavior, please tell me some of the specific behaviors that you considered respectful when you completed your ratings.

For disruptive behavior, please tell me some of the specific behaviors that you considered disruptive when you completed your ratings.

On the days when you completed these ratings, how did you work with students that you rated?

Prompt(s): How did you work with each of the students that you rated? Did you work individually with the student, or did you observe them as part of a group activity?
How did you determine the percentage of time the student exhibited the target behavior? How did you decide on your rating?

Asked of participants who completed the training:

How do you feel about the training?

Prompt(s): Was it helpful?
Was it easy to understand?
Would you prefer having the training online or in person as an in-service?
How would you improve the training?
How did the training affect how you completed your ratings?

Asked of participants who did not receive the training:

Would some sort of training be helpful to you for completing the DBR ratings?

How would a DBR training program be helpful to you?

Ideally, what would a training program for using DBRs consist of?

Finally, how likely would you be to utilize DBRs in the future?

Are there any changes that would increase your likelihood of using DBRs?

Thank you for participating in this interview and for the information that you provided.

Appendix C

Generalizability Theory Statistical Procedures

Organization and Scope

The following appendix contains a discussion of the statistical procedure used in the present generalizability (G) and dependability (D) studies. This discussion draws heavily from Robert Brennan's 2001 text Generalizability Theory, which contains extensive information regarding issues related to various G and D study designs, as well as the statistical procedures used in generalizability theory (G theory). Although several procedures are available to estimate variance in G and D studies, this discussion focuses on the expected mean square (EMS) procedure because the EMS procedure is the statistical procedure used in GENOVA, a computer program often used to estimate variance components for G studies.

In addition, the present discussion also briefly covers the analogous-ANOVA procedure (Henderson's method 1) because the present study uses UrGENOVA (Brennan, 2001) to estimate G study random effects variance components due to missing data. UrGENOVA (Brennan, 2001) is a software program designed to estimate random effects variance components for G study designs that are unbalanced due to nesting or missing data. To perform those estimates, UrGENOVA uses the analogous-ANOVA procedure. Per the recommendation from Brennan (2001), the present study then uses the UrGENOVA G study variance components in D studies with universes of generalization that do not contain missing data and that have fixed facets. The present discussion briefly covers statistical procedures for those D studies as well.

Brennan (2001) also contains information regarding alternative methods for estimating variance, and overall Brennan's book is an excellent resource for gaining a comprehensive understanding of G theory.

Statistical Procedures

The following discussion of the statistical procedures used to estimate variances for the present studies starts with a random effects model containing three facets from the present studies: Students, Raters, and Days. Unlike the present studies, this first random model does not contain any nested facets. First, the present discussion covers the statistical procedure used to estimate variances for a G study using this model, and then the discussion covers the statistical procedure for estimating the variance components in a D study using that same model. In addition, two coefficients used in G theory are introduced, and the statistical procedure used to calculate them is explained. Next, this discussion covers a random model with facets of Students, Raters, Occasions, and days in which the Occasions are nested within the Days. That is the model used for the present G studies. After that, this discussion covers a mixed model for a D Study in which the Occasions facet is nested in Days and both the Occasions and Days facets are fixed, which is the model used for the present D studies.

To begin, it helps to consider the linear model for a G study:

$$X_{\omega} = \mu + \sum \nu_{\alpha} \qquad (1.0)$$

X_{ω} = the observed score

ω = all indices in a design

μ = the grand mean in the universe

α = the indices in the design and their interactions

ν_{α} = the score effect associated with each indice and interaction except μ

$\sum \nu_{\alpha}$ = the sum of the score effects, except (the grand mean), in the design

The above model is a general representation of the linear model for all G study designs. Please note that for a G study only a single score for an

individual or item from the population is under consideration. Mean scores across individuals or items are D study considerations.

For example, please consider a G study for a single student's scores score across a single rater and day in the universe of admissible observations. The linear model for such a design is printed below:

$$X_{srd} = \mu + \nu_s + \nu_{r+} \nu_d + \nu_{sr} + \nu_{sd} + \nu_{rd} + \nu_{srd} \quad (1.1)$$

Table 7 provides a visual representation of how the DBR scores from such a design can be organized for analysis. Figure 3 provides a Venn diagram of how the variance for such a study is divided conceptually.

To understand how G-theory employs equations such as 1.1 to estimate variance associated with each facet and the object of measure, please consider that the linear model printed above is tautological. In other words, the symbols on the right-hand side of the equation represent another way of expressing the symbols on the left-hand side of the equation. As a result, an observed DBR score on a student by a certain rater on a certain day equals the grand mean of those facets in the population and universe plus the score effects associated with each facet and interaction. The grand mean in a linear model for a G study is the mean of the condition, individual, or item across each facet in the population and universe. For linear model 1.1., the grand mean in the population and universe for 1.1 is expected to equal an observed DBR score for a particular student by a particular rater on a particular day.

$$\mu_\alpha = \mathbf{E_s E_r E_d} X_{srd} \quad (1.2)$$

The symbol "**E**" expresses the idea of an expected value assuming that the condition, individual, or item is randomly selected from the population or universe. Underneath the symbol "**E**" a subscript designates the facet from which the individual, item, or condition is selected. In words, the expected value of the mean of an observed DBR score across a student, rater, and day randomly selected from the universe or population is the grand mean of DBR

scores across all of the facets in the universe of admissible observations: students, raters, and days. The universe or population mean for each variance component or score effect is defined below:

$$\mu_{\alpha} = \mathbf{E}_{\acute{\alpha}} X_{\omega} \quad (1.3)$$

X_{ω} = The observed score effect or variance component

α = The index or indices associated with the score effect

μ_{α} = The universe or population mean associated with the score effect

$\acute{\alpha}$ = The set of all indices in ω not contained in α

For example, the population mean for students in the design under consideration is:

$$\mu_{s} = \mathbf{E}_{r}\mathbf{E}_{d} X_{srd} \quad (1.4)$$

In words, the mean DBR score of all students in the population of students rated by DBRs is expected to equal an observed DBR score rated by a particular rater on a particular day in which each condition is randomly selected from the universe of raters and days. For raters, the mean DBR score by all raters in the universe of raters is expected to equal an observed DBR score on a particular student on a particular day.

$$\mu_{r} = \mathbf{E}_{s}\mathbf{E}_{d} X_{srd} \quad (1.5)$$

The universe means for the other facets are defined in the same way. As the preceding definitions suggest, the expected value for each score effect or variance component is zero; otherwise, the equation would not be tautological.

$$\mathbf{E}\nu_{\alpha} = 0 \quad (1.6)$$

As a result, the expected value of a score effect or variance component over any of its subscripts is zero. For example, the expected value of the interaction between students and days for any student randomly drawn from the population of students is zero.

$$\mathbf{E}\nu_{sd} = 0 \quad (1.7)$$

For an infinite population or universe, which the present random model assumes, the universe mean, population means, and the grand mean are unobserved. As a result, these values need to be estimated using data collected through G studies.

To estimate the score effect and variance components in G theory using observed scores collected in G studies, G theory expresses score effects in terms of mean scores. Brennan (2001) provides an algorithm for expressing score effects in terms of mean scores, which is reprinted below.

Step 0: Start with the universe or population mean score μ_α

Step 1: *Minus* the *mean scores* for components that consist of the s nesting indices and m-2 of the primary indices

Step 2: *Plus* the *mean scores* for components that consist of the s nesting indices and m-j of the primary indices

.

.

.

Step j: *Plus* (if j is even) or *Minus* (if j is odd) the *mean scores* for components that consist of the *s* nesting indices and *m-j* of the primary indices

In words, first the algorithm starts with the universe or population mean of the component corresponding to the score effect, μ_α. Then, in Step 1, the mean scores that consist of m-1 of the primary indices are subtracted from the universe or population mean of the variance component or score effect to be expressed. Because linear model 1.1 does not contain any nesting indices, the explanation for the portion of the algorithm that concerns nesting indexes will be covered later. In Step 2, the mean scores for components consisting of m-2 primary indices are then added. After that, the mean scores for components that consist of m-j of the primary indices are added if the number of components added in Step 2 is even or subtracted if the number

of components added in Step 2 is odd. In this explanation, the term *j* refers to the number of components added in Step 2, and the term *m* refers to refers to the number of variance components. For example, for the score effect V_{srd} in linear model (1.1):

Step 0: Start with μ_{srd}

Step 1: Subtract μ_{sr}, μ_{sd}, and μ_{rd} because M=3 and 3-1 = 2.

Step 2: Add μ_{s} , μ_{r}, and μ_{d} because M =3 and 3-2 = 1.

Step J: Subtract μ because M=3, 3-3 = 0, and J is odd.

As a result, the score effect V_{srd} in linear model 1.1 is expressed using mean scores as:

$$V_{srd} = \mu_{srd} - \mu_{sr} - \mu_{sd} - \mu_{rd} + \mu_{s} + \mu_{r} + \mu_{d} - \mu \quad (1.8)$$

Once variance components are expressed in terms of mean scores, the variances can be estimated through ANOVA computations using data gathered through a G study. To begin this process, the means computed using data gathered through the G study must be distinguished from the universe and population mean scores theorized to exist assuming repeated independent random sampling. These unobserved scores have observed score analogues computed using data gathered through the G study. The symbol $\overline{X}_{\alpha}$ denotes the observed mean score for a variance component calculated from the data gathered in a G-study. $\overline{X}_{\alpha}$ is computed using the equation:

$$\overline{X}_{\alpha} = \frac{1}{\pi(\acute{\alpha})} \sum_{\acute{\alpha}} X_{\omega} \quad (1.9)$$

As in equation 1.3, ά refers to the indices in ω that are not included in α. The summation $\sum_{\acute{\alpha}} X_{\omega}$ runs from 1 to the sample size n associated with the index. πά is 1 if the indices are ω; otherwise, πά is the product of the sample sizes for all of the facets in ά. For example, for linear model (1.1):

$$\overline{X}_{r} = \frac{1}{n_s n_d} \sum_{s=1}^{n_s} \sum_{d=1}^{n_d} X_{srd} \quad (2.0)$$

and

$$\bar{X}_{sr} = \frac{1}{n_d}\sum_{d=1}^{n_d} X_{srd} \quad (2.2)$$

Finally, the grand mean $\bar{X}$ is:

$$\bar{X} = \frac{1}{n_s n_r n_d}\sum_{s=1}^{n_p}\sum_{r=1}^{n_r}\sum_{d=1}^{n_d} X_{srd} \quad (2.3)$$

Using this notation, the observed score analogues for the score effects in 1.1 expressed as mean scores can be written as:

$$\bar{X}_{srd} = \bar{X} + (\bar{X}_s - \bar{X}) + (\bar{X}_r - \bar{X}) + (\bar{X}_d - \bar{X}) + (\bar{X}_{sr} - \bar{X}_s - \bar{X}_r + \bar{X}) + (\bar{X}_{sd} - \bar{X}_s - \bar{X}_d + \bar{X}) + (\bar{X}_{rd} - \bar{X}_r - \bar{X}_d + \bar{X}) + (\bar{X}_{srd} - \bar{X}_{sr} - \bar{X}_{sd} - \bar{X}_{rd} + \bar{X}_s + \bar{X}_r + \bar{X}_d - \bar{X}) \quad (2.4)$$

In addition, the linear model for the observed score analogue is:

$$\bar{X} = \bar{X} + x_s + x_r + x_d + x_{sr} + x_{sd} + x_{rd} + x_{srd} \quad (2.5)$$

The lowercase x_α is used to denote the observed score analogue for score effect ν_α.

To estimate the variance associated with each effect in a study in G theory, first the sums of squares and mean squares are computed for each facet and the object of measure. The sum of squares for any component is:

$$SS_{(\alpha)} = \text{πά}\sum x^2{}_\alpha \quad (2.6)$$

In addition, the total sum of squares is:

$$SS_{(tot)} = \sum_\omega (X_\omega - \bar{X})^2 \quad (2.7)$$

For equation 2.6, the summation is taken over all facets in α. For equation 2.7, the sum of squares is taken over all of the indices in the design. For example, the sum of squares for the x_r component in equation 2.5 would be:

$$SS_r = n_s n_d \sum_{r=1}^{n_r} x^2{}_r \quad (2.8)$$

or

$$SS_r = n_s n_d \sum_{r=1}^{n_r} (\bar{X}_r - \bar{X})^2 \quad (2.9)$$

and the sum of squares for the X_{sr} component would be:

$$SS_{sr} = n_d \sum_s \sum_r x_{sr}^2 \quad (3.0)$$

or

$$SS_{sr} = n_d \sum_s \sum_r (\bar{X}_{sr} - \bar{X}_r + \bar{X})^2 \quad (3.1)$$

The total sum of squares would be:

$$SS_{tot} = \sum_s \sum_r \sum_d (\bar{X}_{srd} - \bar{X})^2 \quad (3.2)$$

Using the above equations to calculate the sum of squares for a component can be complicated and time consuming. Brennan (2001) offers a simpler procedure for calculating the sum of squares for components using the total of the sum of squared mean scores of each level or condition of a component. For a component, α, this sum of squared mean scores is:

$$T_{(\alpha)} = \pi(\acute{\alpha}) \sum_{\alpha} \bar{X}^2{}_{\alpha} \quad (3.3)$$

Unlike equation 2.5, which involve x_α terms that symbolize an individual's score over the indices in α, T(α) terms involve total scores for each level of each facet or the object of measure, denoted with the capital $\bar{X}_\alpha$, across the indices in α. In addition, the sum of squared mean scores for the grand mean would be:

$$T_{(\mu)} = \pi(\omega) \bar{X}^2 \quad (3.4)$$

Using "T" terms, the sum of squares can be expressed by replacing the v_α, $\mu_{\alpha,\ and}$ μ terms used in the algorithm provided by Brennan (2001) with the terms $SS_{(\alpha)}$, $T_{(\alpha)}$, and $T_{(\mu)}$. For example, SS_r for equation 1.1 would be:

$$SS_r = T_{(r)} - T_{(\mu)} \quad (3.5)$$

and

$$SS_{(rs)} = T_{(rs)} - T_{(r)} - T_{(s)} + T_{(\mu)} \quad (3.6)$$

Finally,

$$SS_{(tot)} = T_{(pih)} - T_{(\pi)} \quad (3.7)$$

Once the sums of squares are calculated, the mean squares for each variance component are computed using the following equation:

$$MS_{(\alpha)} = \frac{SS_\alpha}{df_\alpha} \quad (3.8)$$

For non-nested effects, the degrees of freedom are the product of (n-1)s for the indices in the effect. N is defined as the G study sample size associated with the index.

Using mean squares, the variance associated with each component in the design can be estimated. To explain this process, it first is necessary to cover how variance components are interpreted in G theory. The interpretation of variance components differs between random and fixed facets, as well as between nested and non-nested facets. Issues associated with models containing nested and fixed facets will be covered in the sections covering nested facets and mixed models.

In a linear model with only random effects, the variance component for each V_α is:

$$\sigma^2(\alpha) = \sigma^2(V_\alpha) = EV^2_\alpha \quad (3.9)$$

In words, $\sigma^2(\alpha)$ is the variance in the population and/or universe of the score effects V_α, and it is the expected value of the square of V_α.

In equation 1.1:

$$\sigma^2(s) = \sigma^2(V_s) = EV^2_s = E(\mu_s - \mu)^2 \quad (4.0)$$

Whenever α is non-nested and not an interaction,

$$\sigma^2(\alpha) = \sigma^2 V_\alpha \quad (4.1)$$

However, $\sigma^2(\alpha)$ does not equal $\sigma^2(\mu_\alpha)$ when α is nested or an interaction. This discussion returns to the issue of nesting later. For the σ^2 associated with an interaction, $\sigma^2(ij)$, where i and j are two facets or indices, $\sigma^2(ij)$ is viewed as the expected value of the square of $(\mu_{ij} - \mu)$ after removing the effects of i and j. For example, for the interaction $\sigma^2(rs)$ in linear model 1.1:

$$\sigma^2(rs) = E[(\mu_{rs} - \mu) - (\mu_r - \mu) - (\mu_s - \mu)]^2 \quad (4.2)$$

or

$$\sigma^2(rs) = E[(\mu_{rs} - \mu) - V_r - V_s] \quad (4.3)$$

Also, G study variance components are viewed as providing a decomposition of total variance. In general:

$$\sigma^2(X_\omega) = E(X_w - \mu)^2 = \sum\sigma^2(\alpha) \quad (4.4)$$

For example, for X_{srd} linear model 1.1:

$$\sigma^2(X_{srd}) = \sigma^2(s) + \sigma^2(r) + \sigma^2(d) + \sigma^2(sr) + \sigma^2(sd) + \sigma^2(rd) + \sigma^2(srd) \quad (4.5)$$

and

$$\sigma^2(X_r) = \sigma^2(r) \quad (4.6)$$

In other words, a variance component in equation 4.2 represents the part of the total variance uniquely attributable to that individual component. To estimate these variance components, a statistical procedure in which mean squares are equated to their expected values is used. This procedure is called the ANOVA procedure. As mentioned above, to find the mean square for an individual component, in other words a facet or interaction, the sum of squares for that component is divided by its degrees of freedom. For example, for linear model 1.1, the mean square for the "s" component is:

$$MS_{(s)} = \frac{n_r n_d}{n_d - 1}\sum_s(\bar{X}_s - \bar{X})^2 \quad (4.7)$$

Although conceptually straightforward, expressing variance components as mean squares can involve a substantial amount of algebra. Fortunately, Brennan (2001) offers a simpler procedure using the notational system employed in his book and throughout this discussion. For any variance component B:

$$EMS(B) = \sum\pi(\acute{\alpha})\sigma^2(\alpha) \quad (4.8)$$

In the above equation, the summation is taken over all the α that contain at least all of the indices in B. Also, as in 1.3, π(ά) refers to the product of the sample sizes of all of the facets in the design not included in α.

Using mean squares calculated from the data collected in a G study, estimated variance components, $\hat{\sigma}^2(\alpha)$, are calculated by substituting the mean

squares from the observed scores, MS(B), for the expected mean squares, EMS(B), in equation 4.8

$$MS(B) = \sum\pi(\acute{\alpha})\,\hat{\sigma}^2(\alpha) \quad (4.9)$$

To find $\sigma^2(\alpha)$ for each variance component α, the MS(B) and $\sum\pi(\acute{\alpha})$ terms are replaced with values calculated using data gathered in the G study. The MS(B) terms are replaced with mean squares for each variance component calculated using the scores collected during the G-study, and the $\pi(\acute{\alpha})$ terms are replaced with the number of levels and conditions associated with each facet or index in the G study not included in α. This replacement procedure results in a system of equations containing an equation corresponding to each variance component in the G study design. For example, for the linear model in equation 1.1, the resulting system of equations would be:

$$MS(s) = \hat{\sigma}^2(srd) + n_s\hat{\sigma}^2(rd) + n_r\hat{\sigma}^2(sd) + n_d\hat{\sigma}^2(sr) + n_r n_d\hat{\sigma}^2(s)$$

$$MS(r) = \hat{\sigma}^2(srd) + n_s\hat{\sigma}^2(rd) + n_r\hat{\sigma}^2(sd) + n_d\hat{\sigma}^2(sr) + n_d n_s\hat{\sigma}^2(r)$$

$$MS(d) = \hat{\sigma}^2(srd) + n_s\hat{\sigma}^2(rd) + n_r\hat{\sigma}^2(sd) + n_d\hat{\sigma}^2(sr) + n_s n_r\hat{\sigma}^2(d)$$

$$MS(sr) = \hat{\sigma}^2(srd) + +n_d\hat{\sigma}^2(sr)$$

$$MS(sd) = \hat{\sigma}^2(srd) + n_r\hat{\sigma}^2(sd)$$

$$MS(rd) = \hat{\sigma}^2(srd) + n_s\hat{\sigma}^2(rd)$$

$$MS(srd) = \hat{\sigma}^2(srd) \quad (5.0)$$

To solve the system of equations in 5.0, the mean squares for variance component $\sigma^2(srd)$ are inserted as the estimate $\hat{\sigma}^2(srd)$. From there, the system of equations is solved in reverse order, from the bottom to the top, by inserting the variance estimates, $\hat{\sigma}^2(\alpha)$, from the equations below the equation being solved. These variance estimates are multiplied by the number of levels and conditions, n_α, associated with each facet.

D Studies. In G studies, researchers estimate the variance associated with each facet in their universe of admissible observations. In D studies, researchers estimate the generalizability of an observed score to scores

obtained from an infinite number of randomly sampled measurements under the same measurement circumstances as defined by the researcher. The source of these measurements, or items, is called the universe of generalization. The average score across all items in the universe of generalization is called the universe score. D studies investigate the relationship between observed scores and universe scores to estimate the error associated with a particular measurement.

The terms "items" and "measurements" refer to the specific instance in which a score is gathered. For example, a DBR measurement as defined in 1.1 would consist of a single score from a DBR filled out by a particular rater on a student on a particular day. While G studies concern the mean score of a single condition sampled from the population, in this case students, over a single level of each facet in the universe of admissible observations, D studies concern the mean score of multiple conditions drawn from the population of the object of measure over the levels of the facets in the universe of generalization. For example, for linear model 1.0, the G study concerns a single student's average score over a single level in the raters and days facets. A D study for that model concerns the mean score of multiple students over multiple levels in the Raters and Days facets.

To make the distinction described above, uppercase letters are used for the facets in the universe of generalization. For example, the sxrxd design from the G study discussion is written sXRXD for D studies. The linear model of the D study for this design is:

$$X_{sRD} = \mu + V_s + V_R + V_D + V_{sR} + V_{sD} + V_{sRD} \quad (5.1)$$

Linear model 5.1 is the decomposition of students' sores over n'_R and n'_D levels of the facets raters and days.

The algorithm for defining score effects in terms of population or universe means described before in the context of G studies also is applicable to D studies. However, while G studies concern the universe means

in the universe of admissible observations, D studies concern universe means in the universe of generalization. For example, for the sxRxD design:

$$\mu_s = E_R X s R \quad (5.2)$$

R is considered the set of random facets in the universe of generalization. For the D study associated with linear model 5.1, those facets are the raters rating the students and the days over which the students are rated. In random models, *R* exhausts all of the levels of the facets in the universe of generalization. As a result, for the sXRXD design, the universe score for students is:

$$\mu_s = E_R E_D X_{sRD} \quad (5.3)$$

because *R* = R and D.

In words, students' universe scores in the universe of generalization are their expected scores over the levels of the facets raters and days in the universe of generalization.

D studies use variance estimates, $\sigma^2(\alpha)$, from G studies to estimate variance associate with each facet in a study. The formula for these estimates is provided below:

$$\sigma^2(\bar{\alpha}) = \frac{\sigma^2(\alpha)}{d(\bar{\alpha})} \quad (5.4)$$

$\sigma^2(\bar{\alpha})$ represents the variance component in the D study design. $d(\bar{\alpha})$ is the product of the D study sample sizes $\grave{n}$ for all indices in $\bar{\alpha}$ except the object of measure. The symbols $\grave{n}$ and $\grave{N}$ are used for the sample size of a facet in a D study and the size of a facet in the universe of generalization respectively. The symbol ` distinguishes these terms from the terms n and N, which are used to indicate the sample size of a facet in a G study and the size of a facet in the universe of admissible observations respectively. For example, from equation 4.8, $\sigma^2(R)$ associated with V_R is:

$$\sigma^2(R) = \frac{\sigma^2(r)}{\grave{n}_r} \quad (5.5)$$

and $\sigma^2(sRD)$ associated with VsRD is:

$$\sigma^2(\text{sRD}) = \frac{\sigma 2(\text{srd})}{\grave{n}_r \grave{n}_d} \quad (5.6)$$

For $\sigma^2(\bar{\alpha})$ associated with the object of measure, s, $d(\bar{\alpha}) = 1$.

Researchers can use D studies to estimate the variance associated with each facet in a D study design given different sample sizes. As a result, researchers use D studies to estimate the number of levels of a facet needed to reduce the variance associated with a facet upon which score variance is undesirable. For the sxRXD random design, researchers would use D studies to estimate the number of raters or days needed to reduce measurement error associated with those facets to an acceptable level.

Generalizability theory conceptualizes two different types of error that are frequently estimated in D studies: absolute error variance and relative error variance. Absolute error variance is the error associated with using an examinee's observed mean score as an estimate of his or her universe score. Relative error variance is the error associated with using an examinee's observable deviation score as an estimate of his or her universe deviation score. Absolute error affects decisions related to an external criterion, such as a cut score. Relative error affects decisions related to the rank order of a person taking a test or being rated.

Absolute error variance is notated as Δ_α. The variance associated with Δ_s, $\sigma^2(\Delta_s)$, is determined by summing the $\sigma^2(\bar{\alpha})$ associated with each variance component in the design with the exception of the variance component associated with the object of measure. In 1.1, that component would be V_s: the score effect associated with students. As a result, the absolute error variance for 1.1 is:

$$\sigma^2(\Delta) = \sigma^2(\text{R}) + \sigma^2(\text{D}) + \sigma^2(\text{sR}) + \sigma^2(\text{sD}) + \sigma^2(\text{RD}) + \sigma^2(\text{sRD}) \quad (5.7)$$

Relative error variance involves adding all $\sigma^2(\bar{\alpha})$ that include the object of measure and at least one other index. For equation 4.7, relative error variance is:

$$\sigma^2(\delta) = \sigma^2(sR) + \sigma^2(sD) + \sigma^2(sRD) \quad (5.8)$$

Absolute and relative error variance are employed to calculate two coefficients in generalizability theory that range from 0 to 1: a generalizability coefficient and a phi or dependability coefficient. The generalizability coefficient employs relative error variance, and the phi coefficient employs absolute error variance. The basic equation for the generalizability coefficient for the sxRxD random designs is:

$$P^2 = \frac{\sigma^2(s)}{\sigma^2(s)+\sigma^2(\delta)} \quad (5.9)$$

The generalizability coefficient is interpretable as an approximation of the squared correlation between universe deviation scores and observed deviation scores.

The basic equation for the phi coefficient is:

$$\Phi = \frac{\sigma^2(s)}{\sigma^2(s)+\sigma^2(\Delta)} \quad (6.0)$$

The phi coefficient is interpretable as an approximation of the squared correlation between the universe scores and observed scores.

G studies with nested facets. To this point, the designs described in this discussion contain only non-nested facets. However, in the present studies, the Occasions facets is nested within the Days facet. Designs with nested facets require a few additional considerations. The present discussion will explain these considerations using the design of the present studies: Students crossed with Raters crossed with Occasions nested in Days (sxrxo:d). To start, the linear model for a G study with the occasions nested within the days facet is:

$$X_{sro:d} = \mu + V_s + V_r + V_d + V_{o:d} + V_{sr} + V_{sd} + V_{so:d} + V_{rd} + V_{ro:d} + V_{srd} + V_{sro:d} \quad (6.1)$$

The above linear model expressed the idea that the facet of Occasions is confounded, and therefore not independent of, the facet of Days. As that description suggests, the effect $V_{srdo:d}$ does not appear in linear model 6.1

because an index cannot appear as both a primary and a nesting index in the same effect. Table 8 provides a visual representation of how the DBR scores from such a design can be organized for analysis. Figure 4 provides a Venn diagram of how the variance for such a study is divided conceptually.

Expressing score effects as mean scores in designs with nested indices requires a modified version of the algorithm used for designs without nested indices. In the modified algorithm, also from Brennan 2001, α is a component with m primary indices and s nesting indices. To express the score effect associated with α, v_α, in terms of mean scores:

Step 0: First, start with the overall mean for the effect, μα

Step 1: Subtract the mean scores for components that contains s nesting indices and m-1 of the primary indices

Step 2: Add the mean scores for the components that consist of s nesting indices and m-2 of the primary indices

Step J: If J is even, add, or if J is odd, subtract, the mean scores for components consisting of s nesting indices and m-j of the primary indices.

Step m: The algorithm stops with the mean score containing only s nesting indices. (6.2)

For example, so:d in 5.4 contains one nesting index, d, and two primary indices, s and o. As a result, in Step 1, M_{sd} and $M_{o:d}$ are subtracted from $M_{so:d}$ because these components contain the nesting index and 1 of the primary indices, s and o, in so:d. To verify, please see that M = 2 and 2-1=1. Next, μ_d is added to the results of step 1 because μ_d involves the nesting index d and 2-2 = 0. Because μ_d contains only the s nesting index, d, the algorithm terminates.

Therefore, $V_{so:d}$ expressed as mean scores is:

$$V_{so:d} = \mu_{so:d} - \mu_{sd} - \mu_{o:d} + \mu_d \quad (6.3)$$

$V_{o:d}$ would be

$$V_{o:d} = \mu_{o:d} - \mu_d \quad (6.4)$$

and $V_{sro:d}$ would be

$$V_{sro:d} = \mu_{sro:d} - \mu_{so:d} - \mu_{ro:d} - \mu_{d} + \mu_{o:d} + \mu_{sd} + \mu_{rd} - \mu_d \quad (6.5)$$

The entire linear model in 5.4 expressed as mean scores is:

$$X_{sro:d} = \mu + (\mu_s - \mu) + (\mu_r - \mu) + (\mu_d - \mu) + (\mu_{o:d} - \mu_d) + (\mu_{sr} - \mu_s - \mu_r + \mu) + (\mu_{sd} - \mu_s - \mu_d + \mu) + (\mu_{so:d} - \mu_{sd} - \mu_{o:d} + \mu_d) + (\mu_{rd} - \mu_r - \mu_d + \mu) + (\mu_{ro:d} - \mu_{rd} - \mu_{o:d} + \mu_d) + (\mu_{srd} - \mu_{sr} - \mu_{sd} - \mu_{rd} + \mu_s + \mu_r + \mu_d - \mu) + (\mu_{sro:d} - \mu_{so:d} - \mu_{ro:d} + \mu_{o:d} + \mu_{sd} + \mu_{rd} - \mu) \quad (6.6)$$

As with models that do not contain nested facets, models containing nested facets such as 6.1 can be expressed in terms of observed mean scores:

$$X_{sro:d} = \overline{X} + (\overline{X}_s - \overline{X}) + (\overline{X}_r - \overline{X}) + (\overline{X}_d - \overline{X}) + (\overline{X}_{o:d} - \overline{X}_d) + (\overline{X}_{sr} - \overline{X}_s - \overline{X}_r + \overline{X}) + (\overline{X}_{sd} - \overline{X}_s - \overline{X}_d + \overline{X}) + (\overline{X}_{so:d} - \overline{X}_{sd} - \overline{X}_{o:d} + \overline{X}_d) + (\overline{X}_{rd} - \overline{X}_r - \overline{X} + \overline{X}) + (\overline{X}_{ro:d} - \overline{X}_{rd} - \overline{X}_{o:d} + \overline{X}_d) + (\overline{X}_{srd} - \overline{X}_{sr} - \overline{X}_{sd} - \overline{X}_{rd} + \overline{X}_s + \overline{X}_r + \overline{X}_d - \overline{X}) + (\overline{X}_{sro:d} - \overline{X}_{so:d} - \overline{X}_{ro:d} + \overline{X}_{o:d} + \overline{X}_{sd} + \overline{X}_{rd} - \overline{X}) \quad (6.7)$$

The sum of squares and mean squares for the components in 6.1 are calculated in almost the same way as for models without any nested facets. The only difference is in the calculation of the degrees of freedom for nested effects. For a nested effect, the degrees of freedom are: [product of (sample size-1)s for primary indices] X [product of sample sizes for nesting indices]. For example, for a G study using linear model 6.1 with 3 occasions over 10 days, the degrees of freedom for occasions nested in days would be 2X10 = 20. If the G-study had 10 students, the degrees of freedom for so:d would be 9X20 = 180. $\hat{\sigma}^2(\alpha)$ for the variance components corresponding to the linear model in 6.1 can be calculated using the procedure described in the section covering models without any nesting.

$$MS(s) = \hat{\sigma}^2(sro:d) + n_o\hat{\sigma}^2(srd) + n_r n_o\hat{\sigma}^2(sd) + n_r\hat{\sigma}^2(so:d) + n_o n_d\hat{\sigma}^2(sr) + n_r n_o n_d\,\hat{\sigma}^2(s)$$

$$MS(r) = \hat{\sigma}^2(sro:d) + n_o\hat{\sigma}^2(srd) + n_o n_s\hat{\sigma}^2(rd) + n_s\hat{\sigma}^2(ro:d) + n_o n_d\hat{\sigma}^2(sr) + n_s n_o n_d\,\hat{\sigma}^2(r)$$

$$MS(o:d) = \hat{\sigma}^2(sro:d) + n_s\hat{\sigma}^2(ro:d) + n_r\hat{\sigma}^2(so:d) + n_s n_r\hat{\sigma}^2(o:d)$$

$$\text{MS(d)} = \hat{\sigma}^2(sro{:}d) + n_o\hat{\sigma}^2(srd) + n_o n_s\hat{\sigma}^2(rd) + n_s\hat{\sigma}^2(ro{:}d) + n_o n_r\hat{\sigma}^2(sd) + n_r\hat{\sigma}^2(so{:}d) + n_s n_o n_r\,\hat{\sigma}^2(d) + n_s n_r\hat{\sigma}^2(o{:}d)$$

$$\text{MS(sr)} = \hat{\sigma}^2(sro{:}d) + n_o\hat{\sigma}^2(srd) + n_o n_d\hat{\sigma}^2(sr) +$$

$$\text{MS(so:d)} = \hat{\sigma}^2(sro{:}d) + n_r\hat{\sigma}^2(so{:}d)$$

$$\text{MS(sd)} = \hat{\sigma}^2(sro{:}d) + n_o\hat{\sigma}^2(srd) + n_r\hat{\sigma}^2(so{:}d) + n_o n_r\hat{\sigma}^2(sd)$$

$$\text{MS(ro:d)} = \hat{\sigma}^2(sro{:}d) + n_s\hat{\sigma}^2(ro{:}d)$$

$$\text{MS(rd)} = \hat{\sigma}^2(sro{:}d) + n_o\hat{\sigma}^2(srd) + n_s\hat{\sigma}^2(ro{:}d) + n_o n_s\hat{\sigma}^2(rd)$$

$$\text{MS(srd)} = \hat{\sigma}^2(sro{:}d) + n_o\hat{\sigma}^2(srd)$$

$$\text{MS(sro:d)} = \hat{\sigma}^2(sro{:}d) \qquad (6.8)$$

The above system of equations is solvable using roughly the same method as used for the model with non-nested facets. The only difference is in the above described way degrees of freedom are calculated for nested facets, which is slightly different from non-nested facets.

Analogous ANOVA procedure. Because the present G studies are unbalanced due to the presence of missing data, the present study used UrGENOVA (Brennan, 2001) to estimate variance components for the G studies. UrGENOVA (Brennan, 2001) uses the analogous-ANOVA procedure to estimate random effects variance components for G study designs that are unbalanced due to nesting or missing data because the analogous-ANOVA procedure produces unbiased variance estimates for such designs (Brennan, 2001). The analogous-ANOVA procedure is slightly different from the EMS procedure described in previous sections.

As discussed, in terms of the T-term notation used in Brennan (2001), the T term for an effect α in a balanced design is:

$$T_{(\alpha)} = \pi(\acute{\alpha})\sum_{\alpha}\bar{X}^2{}_{\alpha} \qquad (6.9)$$

However, an analogous T term is defined as:

$$T_{(\alpha)} = \sum_{\alpha} n_{\alpha}\bar{\bar{X}}^2{}_{\alpha} = \sum_{\alpha}\left[\frac{(\sum_{\varepsilon}\chi_{\alpha\varepsilon})}{n_{\alpha}}\right] \qquad (7.0)$$

In the above equation:

$n_{\alpha\dot{}}$ is the total number of observations for a given level of α

∈ is the set of all indices that are not in α

α∈ means all of the indices in the design

The following steps then are used to estimate random effects variance components using the above analogous T terms:

Step 1: Obtain the expected value of each T term with respect to μ^2, the variance components, and their coefficients.

Step 2: Use traditional algebraic procedures (or matrix procedures) to estimate the variance components

In general, the coefficient of μ^2 in the expected value of every T term is

$$k[\mu^2, ET(\alpha)] = n_+ \quad (7.1)$$

In the above equation, n_+ is the total number of observations in the design. Last, in general, the coefficient of $\sigma^2(\alpha)$ in the expected value of the T term for B is

$$k[\sigma^2(\alpha), \boldsymbol{E}T(\beta)] = \sum_{\beta}\left(\sum_{\gamma}\frac{\tilde{n}^2_{\beta\gamma}}{\tilde{n}_{\beta}}\right) \quad (7.2)$$

where

γ is the set of all indices in α that are not in β (if β = μ, then γ = α);

$\tilde{n}_{\beta\gamma}$ is the total number of observations for a given combination of levels of β and γ; and

$\tilde{n}_{\beta}$ is the total number of observations for a given level of β (note that $\tilde{n}_{\beta} = \sum_{\gamma} \tilde{n}_{\beta\gamma}$).

D studies with nested facets. The linear model for a D study with nested facets can be represented in the same way as a D study with non-nested facets. As well, score effects are expressed in terms of mean scores using the same algorithm used for G studies with nested indices. For example, below is the linear model of the score effects in the present set of D studies:

$$X_sRO{:}D = \mu + V_s + V_R + V_D + V_{O:D} + V_{SR} + V_{SD} + V_{SO:D} + V_{RD} + V_{RO:D} + V_{sRD} + V_{sRO:D} \quad (7.3)$$

Expressed as a decomposition of total variance, 6.9 is:

$$\sigma^2(X_{sRO:D}) = \sigma^2(s) + \sigma^2(R) + \sigma^2(D) + \sigma^2(O{:}D) + \sigma^2(sR) + \sigma^2(sD) + \sigma^2(sO{:}D) + \sigma^2(RD) + \sigma^2(RO{:}D) + \sigma^2(sRD) + \sigma^2(sRO{:}D) \quad (7.4)$$

As with D studies that do not contain nested effects, capital letters are used for the facets in the design, excluding the object of measure, to communicate that D study linear models are for the decomposition of the observed mean scores of the object of measurement over sets of conditions in the universe of generalization, rather than single conditions in the universe of admissible observations.

D studies with fixed facets. Although the universes of admissible observations for the present G studies contains random facets, the universe of generalization for the present D studies contains fixed facets. Random effects variance components from G studies can be used in D studies where some of the facets are fixed. The ability to use variance estimates from G studies in D studies with a different structure is considered a strength of G theory.

Because the facets of days and occasions nested in days are fixed in the present D studies, the D studies in the present study are mixed models. For mixed models, an observable mean score can be represented as X_sRF where F is the set of fixed facets and R contains only random facets in the universe of generalization. Because the estimated G study variance components for the present study are for a random model, the following rules can be used to estimate variance not considered error (τ) (i.e., true variance), relative variance (δ), and absolute variance (Δ) for the present mixed model D studies using random effects G study variance components. Those components can be

calculated using equation 5.4, the same equation used to calculate random effects D study components.

Rule 1: $\sigma^2(\tau)$ is the sum of all random effects D study variance components $\sigma^2(\bar{\alpha})$ such that $\bar{\alpha}$ includes τ and does not include any index in *R*

Rule 2: $\sigma^2(\Delta)$ is the sum of all $\alpha^2(\bar{\alpha})$ such that $\bar{\alpha}$ includes at least one of the indices in R; and

Rule 3: $\sigma^2(\delta)$ is the sum of all $\sigma^2(\bar{\alpha})$ such that $\bar{\alpha}$ includes τ and at least one of the indices in R

Applying these rules to the present D-studies:

$$\sigma^2(\Delta) = \sigma^2(R) + \sigma^2(sR) + \sigma^2(RD) + \sigma^2(RO:D) + \sigma^2(sRD) + \sigma^2(sRO:D) \quad (7.5)$$

and

$$\sigma^2(\delta) = \sigma^2(sR) + \sigma^2(sRD) + \sigma^2(sRO:D) \quad (7.6)$$

Given the above rules, the formulas for calculating the generalizability and phi coefficients for the present studies are identical to those used for the random model:

$$P^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)} \quad (7.7)$$

$$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)} \quad (7.8)$$

Therefore, the generalizability and dependability coefficients for the present studies are:

$$P^2 = \frac{\sigma^2(s) + \sigma2(sD) + \sigma2(sO:D)}{\sigma^2(s) + \sigma2(sD) + \sigma2(sO:D) + \sigma2(sR) + \sigma2(sRD) + \sigma2(sRO:D)} \quad (7.9)$$

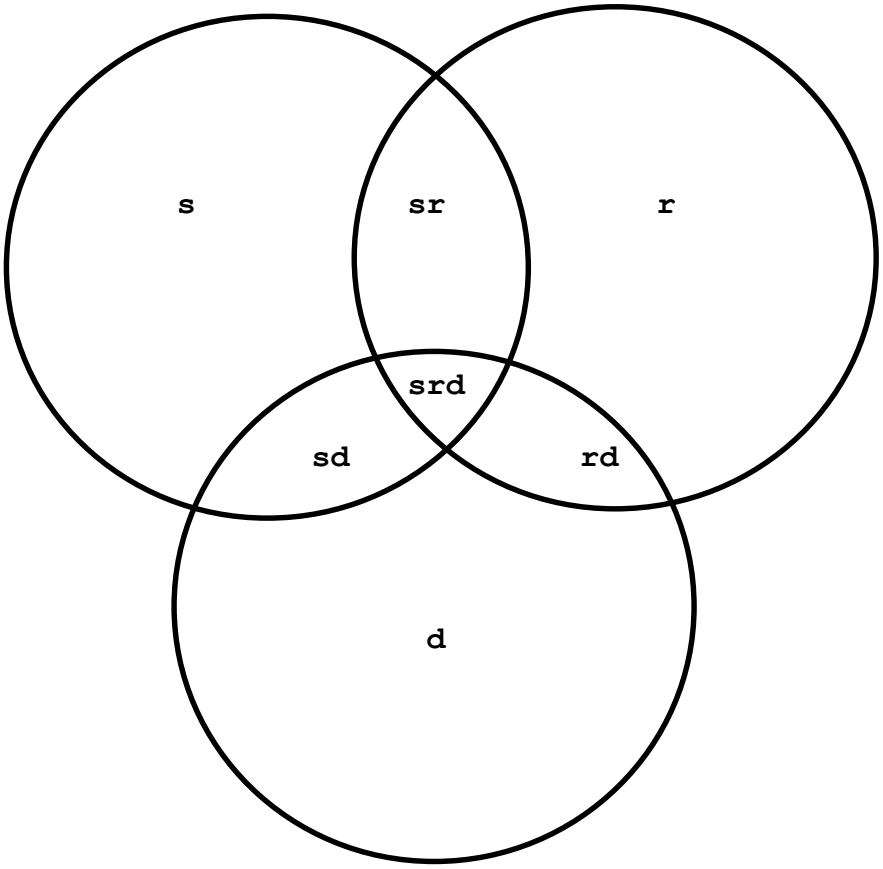
$$\Phi = \frac{\sigma2(s) + \sigma2(sD) + \sigma2(sO:D)}{\sigma2(s) + \sigma2(sD) + \sigma2(sO:D) + \sigma2(R) + \sigma2(sR) + \sigma2(RD) + \sigma2(RO:D) + \sigma2(sRD) + \sigma2(sRO:D)} \quad (8.0)$$

Table 7

Students X Raters X Days Design

	X_{srd}											
	Day 1		Day 2		Day 3		$\bar{X}_{sr}$		$\bar{X}_{sd}$			
Stds	R 1	R 2	R 1	R 2	R 1	R 2	R 1	R 2	Day 1	Day 2	Day 3	$\bar{X}_s$
1	3	6	6	4	8	8	5.6	6.0	4.5	5.0	8.0	5.8
2	5	7	5	5	6	7	5.3	6.3	6.0	5.0	6.5	5.8
3	7	4	7	6	7	9	7.0	6.3	5.5	6.5	8.0	6.6
4	6	7	6	7	5	6	5.6	6.6	6.5	6.5	5.5	6.1
5	9	7	5	8	6	7	6.6	7.3	8.0	6.5	6.5	7.0
6	4	5	3	9	5	5	5.6	6.3	4.5	6.0	5.0	5.2
	$\bar{X}_{rd}$ for Day1		$\bar{X}_{rd}$ for Day2		$\bar{X}_{rd}$ for Day3		$\bar{X}_r$			$\bar{X}_d$		$\bar{X}$
	5.6	6	5.3	6.5	6.2	7	5.9	6.5	5.8	5.9	6.6	6.1

Note: $\bar{X}_\alpha$ represents means of the each level of the indice α across the levels of the other indices not in α. Stds = Students, R 1=Rater 1 and R 2=Rater 2.



Variance Components

Object of Measure	Interactions
Students **(s)**	Students X Days **(sd)**
Facets	Students X Raters **(sr)**
Days **(d)**	Raters X Days **(rd)**
Raters **(r)**	Students X Raters X Days **(srd)**

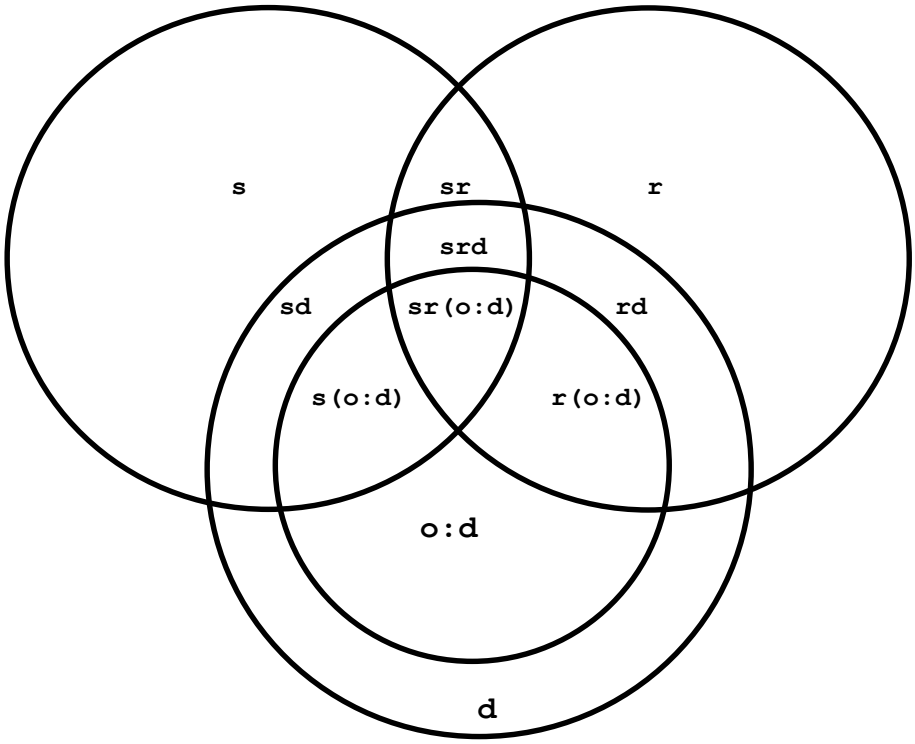
Figure 3. Venn diagram showing a Students X Raters X Days design. Each circle represents a facet or the object of measure. The areas where the circles cross represent interactions. In this design, all of the facets are crossed, and all of the facets are random. Students are the object of measure. Raters and Days are facets.

Table 8

Students x Raters X Occasions Nested in Days Design

	X_{srod}												X_{srd}										
	Day 1												Day 2										
	O1		O2		O3		O4		O5		O6		$\bar{X}_{sd}$		$\bar{X}_{sr}$		$\bar{X}_{so:d}$						
S	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	D1	D2	R1	R2	O:D1	O:D2	O:D3	O:D4	O:D5	O:D6	$\bar{X}_s$
1	5	6	4	5	6	4	4	4	7	2	7	5	5	4.8	5.5	4.3	5.5	4.5	5	4	4.5	6	4.9
2	6	7	5	8	5	3	5	6	5	4	6	4	5.6	5	5.3	5.3	6.5	6.5	4	5.5	4.5	5	5.3
3	4	5	7	5	4	6	4	2	4	5	5	6	5.1	4.3	4.6	4.8	4.5	6	5	3	4.5	5.5	4.7
4	5	6	2	3	4	6	5	4	4	4	5	3	4.3	4.1	4.1	4.3	5.5	2.5	5	5.5	4	4	4.4
5	7	5	7	6	5	4	6	5	3	6	5	7	5.6	5.3	5.5	5.5	6	6.5	4.5	5	4	6	5.3
6	8	4	6	5	6	3	5	4	2	4	6	7	5.3	4.6	5.5	4.5	6	5.5	4.5	4	3	6.5	4.9
	$\bar{X}_{ro:d}$ for O1		$\bar{X}_{ro:d}$ for O2		$\bar{X}_{ro:d}$ for O3		$\bar{X}_{ro:d}$ for O4		$\bar{X}_{ro:d}$ for O5		$\bar{X}_{ro:d}$ for O6		$\bar{X}_{d}$		$\bar{X}_{r}$				$\bar{X}_{o:d}$				$\bar{X}$
	5.8	5.5	5.1	5.3	5	4.3	4.8	4.1	4.1	4.1	5.6	5.3	5.15	4.6	5.0	4.7	5.6	5.25	4.6	4.5	4.1	5.5	4.9
	$\bar{X}_{rd}$ for D1						$\bar{X}_{rd}$ for D2																
	R1			R2			R1			R2													
	5.3			5.0			4.8			4.5													

Note: $\bar{X}_{\alpha}$ represents means of the each level of the indice α across the levels of the other indices not in α.



Variance Components

Object of Measure	Interactions
Students **(s)**	Students X Days **(sd)**
Fixed Facets	Raters X Days **(rd)**
Days **(d)**	Students X Occasions nested in Days **(s(o:d))**
Occasions nested in Days **(o:d)**	Raters X Occasions nested in Days **(r(o:d))**
Random Facets	Students X Raters X Days **(srd)**
Raters (r)	Students X Raters X Occasions nested in Days **(sr(o:d))**

Figure 4. Venn diagram showing a Student X Rater X Occasion nested in Days design. Students (s) are the object of measure, and Raters (r), Days (d), and Occasions nested in Days (o:d) are facets. The facet of Occasions is nested within the facet of Days, and both the facets of Days and Occasions nested within Days are fixed. Each circle represents a facet or the object of measure, and the areas where the circles cross represent interactions. Circles completely enclosed by other circles represent nested facets. The design pictured in this diagram is identical to the one that will be used for the present D studies. The present G studies will be random models (i.e., have all random facets).

Appendix D

Letters of IRB Approval

BALTIMORE CITY
PUBLIC SCHOOLS

Stephanie Rawlings-Blake
Mayor, City of Baltimore

Marnell A. Cooper
Chair, Baltimore City Board of School Commissioners

Gregory E. Thornton, Ed.D.
Chief Executive Officer

May 27, 2016 0000302

Bradley Leposa
625 Hubner Street
Apartment A
Baltimore, MD 21211

Dear Mr. Leposa:

IRB# 0000302

TITLE OF PROPOSAL: *Generalizability of Direct Behavior Ratings by Trained Raters*

This is to notify you of the approval of your project by the Office of Achievement and Accountability (OAA) Institutional Review Board (IRB) for the Protection of Human Subjects. It's the opinion of this Board that you have provided adequate safeguard for the rights and welfare of participants selected for this study. Your proposal seems to be in compliance with OAA's Federal Wide Assurance 00008794 and DHHS Regulations for the Protection of Human Subjects.

Date of Review: 3/24/16

Your approval is valid until 3/23/17. Please note that the assigned IRB number must be displayed on the Informed Consent Form and copies of that form should be submitted to OAA IRB. All research members of this project who will have any interactions with students must be fingerprinted by City Schools Human Capital Office.

This project should be conducted in full compliance with all applicable sections of the IRB Guidelines. The IRB should be notified immediately of any proposed changes. You should also report any unanticipated problems involving risks to participants or others to the IRB. For projects that continue beyond one year from the starting date, the IRB will request continue review and update of the research project. Your study will be due for continue review as indicated above. The investigator must also advise the IRB when this study is completed or discontinued.

If you have any questions, please contact the IRB Chair at (443) 642-4032, or by email at idiibor@bcps.k12.md.us. Thank you for your interest in City Schools.

Respectfully,

Ike Diibor, Ph.D.
IRB Chair

C: Theresa D. Jones, Chief Achievement and Accountability Officer.

PENNSTATE

IRB Program
Office for Research Protections

Vice President for Research
The Pennsylvania State University
205 The 330 Building
University Park, PA 16802

Phone : (814) 865-1775
Fax: (814) 863-8699
Email : orprotections@psu.edu
Web : www.research.psu.edu/orp

EXEMPTION DETERMINATION

Date: February 18, 2016

From: Jodi Mathieu, IRB Analyst

To: Bradley Leposa

Type of Submission:	Initial Study
Title of Study:	Generalizability of Direct Behavior Ratings by Trained Raters
Principal Investigator:	Bradley Leposa
Study ID:	STUDY00004305
Submission ID:	STUDY00004305
Funding:	Not Applicable
Documents Approved:	• DBR Standard Form (1), Category: Data Collection Instrument • Focus Group Interview Script (Part 2) (0.01), Category: Data Collection Instrument • Generalizability of Direct Behavior Ratings Protocol (3), Category: IRB Protocol • DBR Online Training Module Outline (0.01), Category: Other • Focus Group Interview Script (Part 1) (2.01), Category: Other • Screenshots from DBR Online Training Module (0.01), Category: Other

The Office for Research Protections determined that the proposed activity, as described in the above-referenced submission, does not require formal IRB review because the research met the criteria for exempt research according to the policies of this institution and the provisions of applicable federal regulations.

Continuing Progress Reports are **not** required for exempt research. Record of this research determined to be exempt will be maintained for five years from the date of this notification. If your research will continue beyond five years, please contact the Office for Research Protections closer to the determination end date.

Changes to exempt research only need to be submitted to the Office for Research Protections in limited circumstances described in the below-referenced Investigator Manual. If changes are being considered and there are questions about whether IRB review is needed, please contact the Office for Research Protections.

Penn State researchers are required to follow the requirements listed in the Investigator Manual (HRP-103), which can be found by navigating to the IRB Library within CATS IRB (http://irb.psu.edu).

Vitae
Bradley T. Leposa

Education

Penn State State College, PA
M.ED in School Psychology, Fall 2013
Thesis: Parental Employment Status and Children's Academic Achievement

Doctorate in School Psychology, expected graduation spring 2017
Thesis: Generalizability of Direct Behavior Ratings by Trained and Untrained Raters

Credentials and Certificates

Educational Specialist 1, School Psychologist PK-12 Pennsylvania
04/01/2016 – 04/01/2022

SPC 1, School Psychologist Maryland
7/2/2016 – 7/1/2021

Employment

Baltimore City Public Schools Maryland
School Psychologist

Field Experiences in School Psychology

Baltimore City Public Schools Psychological Services Intern Baltimore, MD
Doctoral Level Intern, 2015-2016

Other Experience

Instructor State College, PA
Summer 2015
BBH 411w Research and Applications in Biobehavioral Health

Teaching Assistant, Biobehavioral Health State College, PA
Spring 2013 – Spring 2015

Teaching Assistant, Human Development and Family Studies State College, PA
Spring 2012

Presentation

Clark, T., Crimmins, A., Leposa, B. (2012). *Reading First, or is it?* Presented at the National Association of School Psychologists 2011 Convention, Philadelphia.