5-2020

Strengths-Based Assessment of Students With Emotional Disturbance: Rater Variance Using Generalizability Theory

Kara Alison Loftin

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

STRENGTHS-BASED ASSESSMENT OF STUDENTS WITH EMOTIONAL DISTURBANCE: RATER VARIANCE USING GENERALIZABILITY THEORY

A Dissertation Submitted in Partial Fulfillment
of the Requirement for the Degree of
Doctor of Philosophy

Kara Alison Loftin

College of Education and Behavioral Sciences
School of Special Education

May 2020

This dissertation by Kara Alison Loftin

Entitled *Strengths-Based Assessment of Students with Emotional Disturbance: Rater Variance Using Generalizability Theory*

has been approved as meeting the requirement for the Degree of Doctor of Philosophy in College of Education and Behavioral Sciences, School of Special Education

Accepted by the Doctoral Committee

____
Corey Pierce, Ph.D., Research Advisor

____
Tracy Gershwin, Ph.D., D-BCBA, Committee Member

____
Jason Robinson, Ph.D., D-BCBA, Committee Member

____
Jeri Kraver, Ph.D., Faculty Representative

Date of Dissertation Defense ____

Accepted by the Graduate School

____
Cindy Wesley, Ph.D.
Interim Associate Provost and Dean
Graduate School and International Admissions

ABSTRACT

Loftin, Kara Alison. *Strengths-Based Assessment of Students with Emotional Disturbance: Rater Variance Using Generalizability Theory*. Published Doctor of Philosophy dissertation, University of Northern Colorado, 2020.

Under Part B of the Individuals with Disabilities in Education Improvement Act (IDEA) of 2004, free appropriate public education (FAPE) was guaranteed to all students ages 3 through 21 regardless of ability. For students with emotional disturbance (ED), one of the 13 disability categories served under IDEA, when appropriate special education programming and services were in place early, educators could deescalate the potential for more serious and ingrained behaviors that impacted the course of the student's life beyond childhood. Ensuring FAPE for students with disabilities requires accurate and appropriate identification and progress monitoring data. Data, valid and reliable or not, directly impact a student's special education and related services. Data for these purposes are collected in a variety of ways; one of which is through the administration of behavior rating scales that assess social, emotional, and behavioral domains of students. Typically, assessments take a deficit-based approach; educators need to know where the problem lies. However, a strengths-based assessment approach eases associated stigma, leverages a student's competencies, and might improve parent and school partnerships and communication. This non-experimental study examined rater variance utilizing a strengths-based assessment tool for progress monitoring through G theory, a statistical method to evaluate the dependability of data. A total of 25 middle and high school aged students were rated by three types of raters (parent, teacher, and student self-report) using the Behavioral and Emotional Rating Scale (Epstein & Pierce, in press). Results from the study indicated that when evaluating information from all three rater types together, the student rater significantly impacted the overall understanding of the needs of the students and therefore could impact service implementation and goal development for the student. Results also showed disagreement among rater types based on gender alone. Finally, certain student characteristics were at least partly responsible for some of the raters' inconsistencies in scoring. Results have implications for the development of assessment protocols for raters.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The federal provision of free appropriate public education (FAPE) under the Individuals with Disabilities in Education Improvement Act of 2004 (IDEA) relied upon irrefutable data (Tanner, Eklund, Kilgus, & Johnson, 2018). Inaccurate or inappropriate information may influence a student's access to special education supports and related services. Essential to following the federal mandate under IDEA (2004), one area where accurate data are crucial is in the appropriate identification of students at risk for emotional or behavioral disorders (EBD) who might qualify for special education through IDEA under the category of emotional disturbance (ED).

To provide free appropriate public education for students with disabilities, data from universal screening, on-going progress monitoring, diagnostics, and formative and summative assessments are collected to inform school-based decisions, drive school-wide proactive and preventative programming, and provide the basis for continued evaluation of programming and services (Smith-Millman et al., 2017). Furthermore, data collection, reporting, and response played a key role in the implementation of the Every Student Succeeds Act (ESSA, 2015), the most recent reauthorization of the Elementary and Secondary Education Act of 1965 that replaced the No Child Left Behind Act of 2002 (cited in Yell, 2012). All these mandates and laws sought to improve student outcomes through the use of data to inform school-based decision-making.

To increase information on student performance, school-wide support structures that incorporated both academic and behavioral components such as a multi-tiered system of support (MTSS) or response to intervention (RtI) were implemented. For example, RtI and MTSS are systems that along with assessments in the academic domain, emphasize and rely upon the implementation of screening for emotional and behavioral disorders and at-risk behaviors (Splett et al., 2018). The aim of RtI or MTSS is for schools to implement systematic screening procedures; high quality, evidence-based instruction; continuous progress monitoring; and multiple tiers of progressively more intense instruction (Office of Special Education Programs [OSEP], 2008, 2011). A key source of school-based data in MTSS or RtI was derived from the implementation of behavior rating scales (Tanner et al., 2018). It is important to note that while mentioned in the learning disability identification section of IDEA (2004), these support structures were not specifically referenced by name in federal or mandates. Regardless, these school-wide systems blossomed in response to educators' need to provide the full letter of the law under IDEA.

Background of the Study

As a means to collect data, behavior rating scales might be used for universal screening, progress monitoring assessments, research, and diagnostic purposes. Behavior rating scales are made up of scores or scales derived from a set of aggregated observations that represent the measurement of a latent trait, characteristic, or behavior (Wu, Tam, & Jen, 2016). The constructs measured in behavior rating scales identify both external and internal behaviors (Tanner et al., 2018). By assessing the entire student population, behavior rating scales that have reached technical adequacy through predictive validity and intended as universal screens are the first step to address disproportionality and reduce potential teacher-related biases (Raines, Dever, Kamphaus, & Roach, 2012). Once a student has been identified as at risk for behavioral problems, follow-up assessment tools are needed. A behavior rating scale designed and validated as a diagnostic instrument is used for this purpose.

Behavior rating scales might also be used as progress monitoring tools. To monitor progress, the following questions need to be answered: (a) What measures are needed? (b) How many measures are sufficient? (c) How often should progress monitoring take place? (d) How is responsiveness or non-responsiveness to intervention evaluated and determined? (e) What components of the intervention's effectiveness and efficiency are tied to the information derived from the progress monitoring tool? and (f) Do the behavioral indicators gleaned from the progress monitoring tool give adequate and appropriate information to inform monitoring and evaluation of the intervention (Chafouleas, Riley-Tillman, & Christ, 2009)? In addition to these questions, similar to its use as a universal screen and as a tool for progress monitoring, feasibility concerns arise. Behavior rating scales tend to have a greater number of items, making the length of time needed for informants to complete the assessment cumbersome. Furthermore, the cost for delivering the behavior rating scale on a frequent basis would be prohibitive in many schools. Regardless of these constraints, assessment and evaluation are integral in a data-based, decision making model. The data serve to inform specific practices leading to the delivery of targeted intervention for students with greater need in addition to preventative and proactive programming for all students.

Determining the assessment approach (strengths-based or deficit-based) that provides the most dependable, accurate, and appropriate data is critical to the determination of special education and related services for students with ED. Diagnostic assessment for emotional disturbance is typically based on the identification of the student's deficits (Epstein, 2000). The use of a deficit-based assessment model may not be the result of being drawn to what is familiar (e.g., students are referred for evaluation due to deficits rather than strengths) but simply because the data gathered are more appropriate for diagnostic considerations. For example, the Scales for Assessing Emotional Disturbance (Epstein, Cullinan, Pierce, Huscroft-D'Angelo, & Wery, in press)—one of the most readily available and commonly used diagnostic tools for the evaluation of students with at-risk behaviors who might qualify under the IDEA (2004) category ED—along with other deficit-based scales were designed to meet the IDEA identification criteria and requirements that focus on student deficits; therefore, it is natural the assessment would be deficit-based as well. However, both strengths-based assessment and deficit-based assessment approaches might provide comparable levels of accurate data.

Strengths-based diagnostic assessment for ED is argued to be more effective than deficit-based assessment as it allows professionals to "identify and build on the existing strengths and skills that the child and family present" (Epstein, Rudolph, & Epstein, 2000, p. 50). Additionally, along with the provision of robust information, a strengths-based assessment approach was found to be more readily accepted by various stakeholders in the progress monitoring, diagnostic, and service implementation process (Epstein, 2000). Using a strengths-based assessment allows educators to understand the strengths, skills, and competencies of the student and provide information for the student's individualized education program (IEP) and for transition planning (Duppong Hurley, Lambert, Epstein, & Stevens, 2015). Strengths-based assessment is also thought to stimulate more parental participation in the IEP process and might curb the natural tendency to find someone or something to blame for the student's emotional and behavioral challenges (e.g., parents blaming schools and schools blaming parents; Epstein & Sharma, 1998). One strengths-based assessment tool is the Behavioral and Emotional Rating Scale (BERS-3; Epstein & Pierce, in press), which is used for progress monitoring purposes, decisions that directly impact the intervention, and services chosen to meet the needs of students qualifying for special education and related services under IDEA (2004), in particular those with a diagnosis of ED. Investigating any systematic variation in data across rater types, i.e., the student rating themselves, parent raters and teacher raters, using the BERS-3, could lead to not only the increased use of a strengths-based diagnostic approach with students identified with emotional disturbance but also a more efficient measurement condition using strengths-based behavior rating scales moving forward. An important path to realize this is through the application of generalizability theory (Brennan, 2010).

Generalizability theory is an alternative theoretical approach to psychometric measurement, providing a more comprehensive lens than the oft applied classical test theory (CTT; Brennan, 1992). Generalizability theory delivers a means for decision makers to "extrapolate the results achieved on a limited sample of test tasks, measured under unique test conditions to a universe of tasks and conditions, from which the specific test set has been drawn more or less arbitrarily" (Bloch & Norman, 2012, p.

960). In other words, evaluating behavior rating scales used for diagnostic purposes through the lens of G theory might result in the most appropriate and cost effective means for assessing the emotional and behavioral domains of students that could be delivered with maximum efficiency, demonstrating adequate measurement competence and ability to accurately predict trait strengths to ultimately provide the most fitting special education and related services to students with diagnosed ED (Bloch & Norman, 2012).

Applying G theory to enhance understanding of the technical psychometric measurement qualities of behavioral rating scales, that is, measures of validity and reliability, could clarify systematic errors or variation. In turn, this leads to increased data accuracy (Kane, 1982). Within G theory's framework is the ability to conduct generalizability studies (g study) and then subsequent decision studies (d study). Decision studies provide valuable information for researchers to determine the ideal or optimal conditions for measurement designs for the most accurate and appropriate data interpretation and application (Brennan, 2010; Vispoel, Morris, & Kilinc, 2018). By reducing or even potentially eliminating possible sources of error, reliability increases along with the validity of the scores of the instrument (Kane, 1982). through reliable assessment data, schools make valid decisions that influence the allocation of resources. Fundamentally, reliability is

> the degree to which a set of measurement values can be repeated under precisely the same measurement conditions, thus reflecting the fundamental question in statistics: 'What would happen with the results if I could do the research over again'? (Sijtsma & Van der Ark, 2015, p. 128)

Understanding and controlling for multiple sources of error variation is a means to comprehend underlying factors impacting outcomes derived from rating scales and, as a result, enhances the accuracy of the information collected for support services, accommodations, and/or IEP (Smith-Millman et al., 2017). Parkes (2000) posited two fundamental issues impacting the utilization of behavior assessments in schools: (a) lack of agreement among items and raters along with other sources of measurement error, such as occasion, leading to difficulties in demonstrating high reliability scores; and (b) the high costs associated with the assessment (Kaurin, Egloff, Stringaris, & Wessa, 2016). Contextual appropriateness, i.e., consideration of service delivery needs, alignment with constructs of interest, and adequacy of norms are also fundamental considerations (Kaurin et al., 2016; Tanner et al., 2018). Finally, by using a strengths-based diagnostic assessment approach, families and the students themselves increase engagement in the education process using strengths and family and school resources to reach their IEP goals (Epstein, 2000).

An outcome goal for the incorporation of behavior assessment in a school setting is for the most reliable data to accrue (Parkes, 2000). While applying G theory to behavior rating scales might not directly impact cost to schools, there is an indirect impact as different factors affecting the data that emerge are applied to increase reliability and decrease error variance. Schools might benefit by gaining more accurate data on fewer occasions, for example. Fewer occasions of administering behavior rating scales might reduce some costs associated with the measure. Generalizability theory provides a means to this end.

Significance of the Study

Previous studies have focused on error variance of behavior rating scale scores administered to mainly elementary–aged students (Mason, Gunersel, & Ney, 2014; Rowe, Curby, & Kim, 2019; Smith-Millman et al., 2017; Splett et al., 2018; Tanner et al., 2018; Wolcott & Williford, 2018). No studies have been conducted thus far examining a behavior rating scale used for diagnostic purposes based on a strengths-based approach to diagnostic assessment. Furthermore, no studies using any type of behavior rating scale and generalizability theory have included a population of students beyond elementary school. Investigating the application of G theory on behavior rating scales across participants ages 11 through 18 years old has expanded the use of this notably confusing, yet exceedingly valuable measurement theory to bolster reliability, potentially reduce feasibility concerns, and most importantly to add to the literature on the efficacy and appropriateness of a strengths-based approach to diagnostic assessment and progress monitoring efforts.

The BERS-3 (Epstein & Pierce, in press) uses a strengths-based approach to measure personal strengths and competencies of students including interpersonal strength, intrapersonal strength, affective strength, involvement with family, career strength, and school functioning. This tool could be used as a proactive measure for student placement in specialized services and to measure response to services. Examining the norming data from this assessment through the application of G theory provides a pathway to possibly increase the accuracy and appropriateness of data collected from this type of measure as well as to further examine the appropriate use of an alternative approach to assessment, the strengths-based approach, specifically for students falling under the IDEA (2004) category of ED. Doing so would enhance the identification of students with ED, have a direct impact on the chosen intervention and services to meet the student's specific needs, and enrich continued support and appropriate special education and related services for students.

Problem Statement

Data collected from behavior rating scales are a proactive approach to addressing student need and crucial to delivering federal mandates for FAPE (Bruhn, Woods-Groves, & Huddle, 2014). However, behavior rating scales are not a panacea. Subsequent decisions derived from data collected using behavior rating scales are only as good as the data obtained. If reliability or validity of scores obtained from the instrument are in question, they no longer serve their purpose to inform decision making (Miller, Crovello, & Chafouleas, 2017). Specific school policies and practices are put in place, leading to the delivery of targeted intervention for students with greater need and preventative and proactive programming for all students; however, systematic variation in data could exist across occasions, raters, and items (Peters, Kranzler, Algina, Smith, & Daunic, 2014; Splett et al., 2018). These differences might result in questionable data. Collected data, whether accurate or inaccurate, directly impact the intervention and services chosen to meet students' needs. Specific to school decision makers, for example, administrators and school psychologists, applying generalizability theory (G theory) to universal screens allows individuals to parse procedural inferences impacting rater, occasion, and item variances. In turn, this might inform the creation of comprehensive and widely adopted standards for the operational procedures involved in the implementation of behavior rating scales for school-based decisions.

Purpose of the Study

The purpose of this non-experimental study was to examine rater variance of a strengths-based assessment using generalizability theory (G theory). Applying G theory to a strengths-based assessment approach by investigating behavioral rating scales designed for use with a population of students at-risk for problems in the social, emotional, behavioral, and mental health domains served to expand the current literature on strengths-based assessment approaches, opening a unique, novel pathway for the use of G theory statistical analysis. Understanding factors affecting students' ratings on teacher-rater forms, student-rater forms, and parent-rater forms resulting in different ratings for the same student might inform (a) rater training, (b) selection of raters, (c) provide information on the importance of an assessment approach with multiple data types (e.g., student work, formal assessment, and teacher interview) or (d) might provide rationale for a strengths-based approach over a deficit-based approach to assessment or vice versa.

Research Questions

Q1 To what extent does rater type (e.g., student, parent, and teacher) explain rater scores derived from a strengths-based behavioral rating scale?

Q2 To what extent are the scores derived from three rater informants on a strengths-based behavioral rating scale reliable for use in absolute or relative decisions?

Definition of Generalizability Theory Terms

Crossed and nested designs. A crossed study design is described as a study in which "all individuals are measured on all levels of all facets," whereas a nested study design is a study in which "not all levels of one facet are paired with the levels of another facet" (Shavelson & Webb, 1991, pp. 601, 610).

Decision study (d study). A d study results in the application of the outcomes of the measure across a universe of generalization (a wider set of situations; Brennan, 2003; Shavelson, Webb, & Rowley, 1989). This measurement procedure uses the same design over several separate studies and takes place after the g study is conducted (Kane, 1982).

Dependability coefficient or index of dependability (d coefficient). The dependability coefficient or the absolute coefficient is used for absolute decisions or criterion referenced decisions. It “equals the proportion of person-score variance relative to all effects that influence relative person ordering, thus excluding all effects the relative person ordering does not depend on and that define the scope of the generalization” (Sijtsma & Van der Ark, 2015, p. 133).

Facet. A measurement condition, e.g., rater, item, and occasion (Brennan, 2003).

Fixed and random facets. *“*Fixed facets contribute to the variance of interest (true variance)” and “replicate the conditions of the original study” (Bloch & Norman, 2012, p. 967). Random facets “contribute to the error variance” and are a “sample of a universe of possible allowed conditions” (Bloch & Norman, 2012, p. 967).

Generalizability. This broad, flexible term replaces the term, reliability. “Instead of asking how accurately observed scores reflect their corresponding true score, GT asks how accurately observed scores permit us to generalize about a person’s behavior in a defined universe of situations” (Shavelson et al., 1989, p. 922). Bloch and Norman (2012) listed the following substitutions for generalizability over reliability: (a) inter-rater reliability becomes “To what extent can we generalize these exam scores across raters?”; (b) test-retest reliability becomes

"To what extent can we generalize these scores across occasions?"; (c) "To what extent can we generalize these scores across both occasions and raters?"; and finally (d) "What are the most likely sources of error in this particular measurement situation?" (p. 967).

Generalizability coefficient (g coefficient). The generalizability coefficient, which is g theory's reliability coefficient, "equals the proportion of person-score variance relative to all effects that influence relative person ordering, thus excluding all effects that the relative person ordering does not depend on and that define the scope of the generalization" (Sijtsma & Van der Ark, 2015, p. 133). The g coefficient is used in relative (e.g., college admissions) or norm-referenced decisions (e.g., driver's license).

Generalizability theory. Provides a "framework for examining the dependability of behavioral measurements" (Shavelson et al., 1989, p. 922). Unlike classical testing, which is not able to differentiate the source of the error(s), G theory is a statistical method that concurrently examines multiple sources of error and their interactions, providing estimates of dependability and standard error of the mean (SEM) to identify "optimal measurement procedures through which assessment data may be collected" for relative and absolute decisions (Fan & Hansmann, 2015; Frey, 2018a; Tanner et al., 2018, p. 4). Generalizability theory "pinpoints the sources of measurement error, disentangles them, and estimates each one" (Webb, Shavelson, & Haertel, 2006, p. 93).

Generalizability study (g study). Used to analyze as many sources as possible for variance to calculate the reliability or dependability of the measurement (Kane, 1982). A g study is composed of a universe of admissible observations (Brennan, 2003).

Object of measurement (Facet of differentiation). Typically, the person whom the rater is assessing. However, object of measurement is not always a person (Brennan, 2003). For this study, the object of measurement is a person.

Summary

Irrefutable data are imperative in the appropriate identification of students at risk for emotional or behavioral disorders who might qualify for special education through IDEA (2004) under the category of ED. Furthermore, to follow the federal mandate under IDEA (2004) Part B, irrefutable data are necessary to continue to provide the most effective and appropriate support for students already receiving special education and related services under the category of ED. Using G theory, rater variance between student, teacher, and parent on the BERS-3 (Epstein & Pierce, in press), a strengths-based assessment, was examined to add to the body of knowledge on defensible, efficient, repeatable, and flexible data used for the determination or continuation of special education and related services for students with ED (Parker et al., 2012).

CHAPTER II

REVIEW OF LITERATURE

According to the most recent report issued from the U.S. Department of Education (2018), qualification under IDEA Part B of students ages 3 through 21 attending American schools to receive special education and related services was approximately 14% or 6.7 million. About 5% (*n* = 334,997) were eligible for special education services within the category of emotional disturbance (ED; U.S. Department of Education, 2017). IDEA (2004) operationally defined ED as follows:

> (4)
>
> (i) Emotional disturbance means a condition exhibiting one or more of the following characteristics over a long period of time and to a marked degree that adversely affects a child's educational performance:
>
> > (A) An inability to learn that cannot be explained by intellectual, sensory, or health factors.
> >
> > (B) An inability to build or maintain satisfactory interpersonal relationships with peers and teachers.
> >
> > (C) Inappropriate types of behavior or feelings under normal circumstances.
> >
> > (D) A general pervasive mood of unhappiness or depression.
> >
> > (E) A tendency to develop physical symptoms or fears associated with personal or school problems.

> (ii) Emotional disturbance includes schizophrenia. The term does not apply to children who are socially maladjusted, unless it is determined that they have an emotional disturbance under paragraph (c)(4)(i) of this section. (34 CFR Sec. 300.8 (c)(4))

For those students who struggle in school due to their behavior, an accurate diagnosis followed by educational support and related services is paramount as these students have a high rate of behavioral problems persisting into adulthood, tend to have decreased graduation rates, and are impacted over the long-term, affecting work and family variables (Gage et al., 2010; Mitchell, Kern, & Conroy, 2019). Students within the ED category of IDEA Part B have the highest drop-out rates of any of the 13 qualifying disability categories with rates of 30% or higher being reported (U.S. Department of Education, 2018). Additionally, students with ED have the highest rates of in-school and out-of-school suspension and expulsion for more than 10 consecutive days (U.S. Department of Education, 2018). Finally, ED is among the top three disability categories served under IDEA Part B with students who are unilaterally placed in an alternate educational setting such as self-contained classrooms or residential treatment centers (Mathur & Jolivette, 2012; U.S. Department of Education, 2018). It is imperative to identify and provide support structures early and often for students who fall under the disability category of ED. When appropriate intervention and support services are in place earlier, it is possible to deescalate the potential for much more serious and ingrained behaviors (Pierce, Nordness, Epstein, & Cullinan, 2016).

Alarmingly, most assessments measuring social, emotional, and behavioral traits are not aligned to the federal definition of ED (Cullinan & Epstein, 2013a; Nordness, Epstein, Cullinan, & Pierce, 2014). The disconnect between the federal operational definition of ED and assessment creates a challenge for educators as they must attempt to match assessment information with the federal administrative definition (Cullinan & Epstein, 2013b; Cullinan, Harniss, Epstein, & Ryser, 2001; Nordness et al., 2014). Ultimately, this impacts the identification of qualifying students. There are two essential criteria to consider in the evaluation of a psychometrically sound instrument for assessment of ED: (a) the instrument must be based on the federal operational definition of emotional disturbance and (b) assessment results from these instruments must clearly delineate students at risk and in need of a more comprehensive assessment (Gage et al., 2010; Nordness et al., 2014). According to Pierce et al. (2016), IDEA encouraged schools to use social, emotional, and behavioral assessments to identify students who exhibited problem behaviors but whose condition had not progressed to an emotional or behavioral disorder as a preventative measure along with their use for diagnostic and progress monitoring purposes.

Disproportionality

Underrepresentation of students with ED continues to be a concern. Taking all students attending school in the United States ages 3 through 21-years-old regardless of qualification for IDEA (2004) Part B into account, only 1% of students are identified as ED (Merikangas et al., 2010). This number has remained relatively the same for decades. Yet, estimates as high as 20% prevalence of ED among students is thought to be a more accurate representation of the current school milieu (Merikangas et al., 2010). Emotional disturbance covers a broad range of attributes, encompassing mental, emotional, and behavioral health challenges (Lloyd, Bruhn, Sutherland, & Bradshaw, 2019). The approximation of 20% of the school population as meeting criteria for special education and related services under ED includes students who currently have a mental, emotional, or behavioral disorder or who have had a mental, emotional, or behavioral disorder in the past (Merikangas et al., 2010). This estimate of 20% prevalence among students ages 3 through 21 implies ED is far more widespread than the percentage of students who actually receive special education services under the category of ED through IDEA (Mitchell et al., 2019).

Concerns continue regarding students who might qualify under ED but are underrepresented due to at-risk behaviors that are internal rather than external (Allen, Kilgus, Burns, & Hodgson, 2018; Hanchon & Allen, 2018). It is far easier for educators and school psychologists to identify students who display external at-risk behaviors since many times these behaviors are on full display in the school setting and might directly impact others in the school's ecosystem. In addition to underrepresentation concerns, Hanchon and Allen (2018) underscored the overrepresentation of subgroups of students within the ED category, predominantly (a) African-American students ages 6 through 21, who are 2.14 times more likely to be served under this category than peers; and (b) boys, representing 80% of students served under the ED classification. Among other subgroups of students with disproportionate representation are those from either (a) low socioeconomic status, (b) households with an overall lower level of education, or (c) single-parent households (Becker et al., 2011; Hanchon & Allen, 2018).

Many recommendations have addressed disproportionality, i.e., the under- or over-representation among students of varying demographic backgrounds. Regular universal screening of social, emotional, and behavioral domains with integrated measures for both internalizing and externalizing at-risk behaviors is recommended (Scardamalia, Bentley-Edwards, & Grasty, 2019). Diagnostic assessments should follow, if appropriate, in addition to progress monitoring assessments of these domains (Lloyd et al., 2019; Mitchell et al., 2019). Identification of valid, reliable, and feasible methods to assess classroom and school climate is a means to address disproportionality (Scardamalia et al., 2019). Moreover, data collection measures to assist in the identification of students with mental health concerns who tend to not be satisfactorily identified under the ED criteria including but not limited to mental health disorders associated with trauma that impacts a student's access to education is recommended (Lloyd et al., 2019; Mitchell et al., 2019; Scardamalia et al., 2019).

The Every Student Succeeds Act (2015) along with IDEA (2004) placed a greater emphasis than ever before on accountability through data collection to address disproportionality in addition to providing educational supports and systems to increase a student's access to education (Ruble, McGrew, Wong, & Missall, 2018). Yet, under- and over-representation issues are not the only challenge within the field of ED. Mitchell et al. (2019) addressed three other present concerns including (a) integrated delivery models that allow greater access to related services to enhance the benefits of special education for students with ED; (b) regular universal screening for signs of social, emotional, and/or behavioral at risk-behaviors,; and (c) the use of multi-tiered systems of support that offer preventative interventions to address the social, emotional, and behavioral needs of all students, particularly when they first emerge.

Generalizability Theory and Emotional Disturbance

Through the application of G theory, a conceptual framework for understanding more deeply multiple sources of error variance and its impact on outcomes of the assessment, irrefutable data emerge (Lakes & Hoyt, 2009). Examining the interpretation of scores derived from three rater types (student, parent, and teacher) through the application of G theory using an assessment measure such as the BERS-3 (Epstein & Pierce, in press), whether functioning as a progress monitoring measure or a much-needed measure collected on the road to diagnostic assessment, is a neglected area of study that warrants further investigation (Dowdy & Kim, 2012; Gage et al., 2010). This psychometrically sound and frequently used behavioral assessment tool, coupled with generalizability theory, might provide a pathway to address reliability concerns and to streamline procedural applications by decreasing cost and time constraints and delineating students at risk for internal and external behaviors (Gage et al., 2010; Nordness et al., 2014). Furthermore, using G theory to look more closely at the potential variance based on rater type is necessary as the standard process for selection of teacher and parent to complete these types of scales has not been clearly defined (Ikeda, Neessen, & Witt, 2008). For example, students' general education or homeroom teachers typically act as teacher informants (Pierce et al., 2016). However, what has been lacking is a clear understanding of the delineation of which types of teachers provide the greatest inter-rater reliability or even how administration training impacts rater reliability (Dowdy & Kim, 2012). Information gleaned from this study could inform future assessment practices for the on-going monitoring and provision of services for students identified with ED. Optimal testing conditions might increase the adoption and feasibility of behavioral rating scales in school settings and correspondingly improve the reliability of data collected in order to serve students more appropriately, preventatively, and proactively. These data impact administration, usage, and planning of optimal measurement protocols for behavior rating scales that provide the most accurate and appropriate data for the provision of the federal mandate of a free appropriate public education under IDEA (2004; Fan & Sun, 2014).

Behavior Rating Scales

Behavior rating scales are fundamental to data collection efforts outside of the academic domain. Sources for the initial evaluation of special education services under the category of ED include (a) parent interview, (b) classroom observation, (c) teacher interview, (d) student interview, and (e) normative data from behavior rating scales completed by at least two of the three following individuals: the student, a parent, and/or a teacher (Scardamalia et al., 2019). Assessment through behavior rating scales might close the gap between the initial presentation of at-risk behaviors and subsequent service implementation (Volpe & Briesch, 2016). Evidence-based behavior rating scales marry research and theory so (a) appropriate constructs are selected for a given assessment item, (b) the most ideal instruments and methods are used to inform the decision-making process, (c) utility and accuracy of data are optimized, (d) costs are taken into consideration, and (e) student outcomes derived from the assessment process are maximized (Hunsley & Mash, 2018). As tiered intervention frameworks (e.g., RtI or MTSS) grow in use in a data-based decision-making model, schools, in turn, must align resources to student need identified through the assessment process (Chafouleas et al., 2009). With RtI or MTSS, there are three levels of academic and behavioral support: (a) core instruction provided to all students, (b) core instruction and additional support for some students, and (c) core instruction and intensive support for a few students (Chafouleas et al., 2009). Behavior rating scales, whether functioning as universal screens, progress monitoring measures, or diagnostic tools, are crucial fundamental elements to determine a student's placement in this tiered system.

Rating Scales for Progress Monitoring

Progress monitoring is a method to gather data on a student's progress in relation to age or grade expectations or benchmarks at predetermined times throughout the school year and indicates whether the implemented special education and related services have been effective or not (Wixson & Valencia, 2011). This ongoing, frequent measurement assesses interventions and informs instructional and intervention decisions (Kratochwill, Clements, & Kalymon, 2007). Typically, two types of assessments are used for progress monitoring measures: behavior rating scales and systematic direct behavior observation (Volpe & Briesch, 2016). As an evaluation tool, progress monitoring does not provide adequate information on specific instruction strategies or services to implement on behalf of the student but it does offer valuable information on how the student is or is not improving in the domain of interest or concern (Wixson & Valencia, 2011). It is important to note that when using G theory, progress monitoring assessments are absolute decision studies (Fan & Hansmann, 2015).

Among the deterrents for the utilization of progress monitoring is informant load. Informant load refers to the length of the rating scale or the number of items and the frequency raters or informants are asked to complete the assessment (Volpe, Briesch, & Gadow, 2011). It could also simply refer to the burden placed on educators whose time is exceedingly valuable. To address this, some schools have taken subsets of items from rating scales to create progress monitoring measures (Volpe et al., 2011). However, modifying existing behavior rating scales is not recommended due to the validity and reliability concerns arising from an altered behavior rating scale form. An altered behavior rating scale is no longer evidence-based. For schools who do not adapt behavior rating scales, feasibility could be a deterrent. Many rating scales contain 50 or more items, which might negatively impact the acceptability of the scale as a progress monitoring assessment along with the raters' willingness to participate in the process. Data collection for progress monitoring must happen frequently and systematically for the most benefit (Volpe & Briesch, 2016). Nonetheless, due to their comprehensive nature, behavior rating scales tend to have many items. Reducing the number of items on the scale might prove useful for progress monitoring needs; however, new explorations of validity and reliability measures are required. Generalizability theory could address this challenge through effective and efficient means.

Diagnostic Usage

Qualification for special education and related services under IDEA (2004) is not contingent upon a diagnosis of a disability. Moreover, diagnostic determination is not the purview of special educators but rather school psychologists. However, behavior rating scales might be implemented as one tool to gather information to inform school psychologists regarding diagnostic considerations. The behavior rating scale in this study, the BERS-3 (Epstein & Pierce, in press), has not been explicitly used as a diagnostic tool; however, it functioned as one of several tools used to gather information that might lead to a diagnosis. Diagnostics involve a criterion or absolute decision involving cut-off scores; they tend to have many constructs of interest or items/tasks being measured than other types of assessments (Briesch, Chafouleas, & Johnson, 2016). However, due to the technical adequacy, the high reliability, and validity associated with the outcomes of the scores and the identification of both at-risk internal and external behaviors within the framework of the federally operationalized ED definition, behavior rating scales whose main purpose is for diagnostic consideration are attractive to the practitioner (Briesch et al., 2016). Generalizability theory allows researchers to investigate this type of usage.

Deficit-Based Versus Strengths-Based Diagnostic Assessment

Behavior rating scales, whether functioning as universal screens, progress monitoring measures, or diagnostic assessment tools, might be based on the deficit model of assessment or the strengths-based model of assessment. Similar to the medical model versus the social model of disability, which frames differences found in humans as either impairments and deficits resulting as the consequence of a condition, disease, or trauma with the intent on preventing, or treating and curing the condition (medical model) or alternatively, the belief that these variations in humans were not about impairment or deficit but the result of artificially constructed societal and environmental barriers (social model), the type of assessment approach the diagnostician embraces could result in significantly diverse information with distinctive implications for usage. Deficit-based assessment has evolved to a high level of identification and documentation of the areas considered wrong in the student (Epstein, 2000). Students might be labeled and described by their pathologies, by their problems, and by their deficits such as "oppositional defiant disorder," a label that is at its core deficit-based (Epstein et al., 2000). Deficit-based models of assessment are universally used with a multitude of psychometrically sound instrument tools to choose from, all assessing and increasing understanding of a student’s behavior (Epstein, 2000). For planning and evaluation, a deficit-based model of assessment provides the necessary information. However, Mooney, Epstein, Ryser, and Pierce (2005) argued that in addition, deficit-based models were restrictive and provided a myopic focus from the professionals conducting the assessment. Consequently, the development, implementation, and monitoring of special education and related services might be hindered (Mooney et al., 2005).

A strengths-based assessment is defined as

> the measurement of those emotional and behavioral skills, competencies, and characteristics that create a sense of personal accomplishment; contribute to satisfying relationships with family members, peers, and adults; enhance one's ability to deal with adversity and stress; and promote one's personal, social, and academic development. (Epstein & Sharma, as cited in Epstein, 2000, p. 249)

Four basic principles of strengths-based assessment are (a) every child has strengths that are uniquely theirs, (b) the absence of demonstrating an emotional or behavioral strength is not equivalent to a deficit in those areas—it is an indication of a lack of experiences to master the skill in question, (c) focusing on what the student can do rather than cannot do motivates that student and influences him or her to respond positively, and (d) strengths-based IEP objectives and goals promote personal, social, and academic development (Epstein et al., 2000). The student’s skills, affinities, and family and life histories come to the surface through this assessment approach (Epstein, 2000). Additionally, the unmet needs of the student are uncovered (Epstein, 2004). A strengths-based approach with its emphasis on the strengths, potential, capabilities, competencies, and resources of the student and his or her family provides a platform for scaffolding student needs, informing the development of a treatment plan, individual service plans, transition plans and services, and IEP (Epstein, 2000). Instead of asking "What is wrong with this student?" a strengths-based approach asks, "What is right with this student?" A strengths-based assessment approach investigates beyond the school environment and includes addressing the student's and family's specific needs (Epstein, 2004; Munger, 2000). Life domains and environments such as residential concerns; areas of familial, social, educational, vocational, and health interests; and psychological, legal, and safety-related factors are examined using a strengths-based assessment process (Munger, 2000). Regardless of these differences, both deficit-based and strengths-based assessment approaches are united in purpose, i.e., to identify "the gap between a person's capability and the demands of the environment" (Wehmeyer, 2019, p. 11). Subsequently, these data are used to directly inform instructional and behavioral interventions by providing the most accurate and appropriate special education and related services chosen specifically to meet the individual student's unique needs.

Measures of Individual Differences: Data and Psychometric Properties

A data-driven, decision-making model is essential as schools work toward continuous improvement, meet accountability requirements, hone in on specific school-wide efforts, and capture accurate representations of the effectiveness of interventions and supports (Goldring, Berends, & American Association of School Administrators, 2009). Ergo, the way data are collected, what data are collected, the tools used to collect the data, and how the data are analyzed, namely, the procedures that dictate application and use of the assessment tools, are of utmost importance. If data are deemed less than robust, there is a direct impact on the usefulness of the collected data (von der Embse & Kilgus, 2018). Moreover, there is a direct impact on the interventions and services chosen to meet that child's needs.

A barometer for reliable and valid data is defensibility, flexibility, efficiency, and repeatability (Chafouleas et al., 2009). Data that are defensible are standardized, accurate, reliable, and valid (Chafouleas et al., 2009; Parker, Vannest, Davis, & Clemens, 2012). Flexible data are data that could be used for a variety of means, behaviors, purposes, and contexts (Chafouleas et al., 2009). Flexible data include data measures used for identification of behavior problems, allowing for the magnitude of the problem to be evaluated. Flexible data also are data employed as measures for monitoring progress and interventions, evaluation of problem solutions, and for diagnostic and classification decisions (Chafouleas et al., 2009). It is important to acknowledge that the use of tools found to be psychometrically sound (reliable and valid) is a universally known truth among researchers. However, equally important is the systematic evaluation, identification, and promotion of evidence-based assessment procedures that guide how data are collected and evaluated (von der Embse & Kilgus, 2018). Efficiency in reference to data means ratings are completed by people who are naturally part of the student's ecosystem and the measures take relatively brief periods of time to complete (Chafouleas et al., 2009). Efficiency can also be viewed as feasibility in terms of time and financial resource requirements (Webb et al., 2006). Finally, repeatable refers to the appropriateness of the measure to be used multiple times within and across occasions (Chafouleas et al., 2009). Viewing behavior measurements through generalizability theory rather than classical test theory provides a means for school decision makers to maximize the efficiency, defensibility, repeatability, and flexibility of the data derived from the behavior rating scale all while keeping at the forefront the purpose of these endeavors, which is to increase a student's access to education through free appropriate public education. Using a strengths-based assessment tool in this study provided invaluable information that bolstered the defensibility, flexibility, efficiency, and repeatability of the data.

Behavior and Psychometric Assessments

Behavioral assessments, such as behavior rating scales, rely on the assumption of steady state behavior (Shavelson et al., 1989). Nevertheless, behavior is not a person's "property or attribute…[behavior] happens when there is an interactive condition between a [person] and its surroundings" (Johnston & Pennypacker, cited in Cooper, Heron, & Heward, 2007, p. 25). Behavior is determined by the environmental context (Cooper et al., 2007) and shaped through antecedent and consequence variables in one's environment. Therefore, behavior is not an inherent, fixed trait. Behavior is malleable and specific to the environment from which it springs; because of its nature, measurements of behavior contain errors.

Behavioral assessments are intended to be repeatable measures comparing the individual to themselves as well as comparing the individual to others at different developmental points and different times. Traditionally, classical testing theory is used to evaluate psychometric assessments. However, for behavioral assessments in an applied setting with a multitude of potential factors influencing the data, the environmentally shaped (e.g., unstable) characteristic of behavior is better served through the application of generalizability theory that provides a sharper picture of sources of error in addition to an understanding of the extent the assessment data are generalizable. Table 1 provides a brief explanation of unstructured and structured measurement context setting outcome expectations for studies employing G theory.

Table 1

Expected Generalizability Coefficient: Levels of Structure in Measurement Context and Context Dependence of Behavior

		Behavior	
		Context dependent	Context independent
Measurement	Unstructured	Low g	High g
context	Structured	High g	High g

Source: Bruckner, Yoder, and McWilliam (2006).

Measurement Error and Data

All measures of human behaviors or traits involve some level of error (Drost, 2011). Error refers to variations in data due to measurement conditions (Brennan, 2010). This error can be systematic or unsystematic, known or unknown, and fixed or variable (Brennan, 1992). Through the application of the statistical framework generalizability theory, the dependability or reliability of behavior measurements can be evaluated for sources of measurement error (Webb & Shavelson, 2005). Error can be examined, estimated, and disentangled to inform procedural considerations, testing conditions, and feasibility concerns arising from the data collection process and the data itself (Webb & Shavelson, 2005; Wu et al., 2016). Advances in statistical procedures beyond the traditionally employed CTT on measurement of individual traits and behaviors are a means to address previously mentioned concerns regarding error, feasibility, approach to assessment (strengths-based and deficit-based), and identification of students who might qualify for special education and related services under the category of ED (Parker et al., 2012; von der Embse & Kilgus, 2018). In a data-based decision-making model, sources of error must be reduced to ensure the highest levels of reliability and validity—the main benchmarks for assessment and evaluation (Mushquash & O'Connor, 2006).

Reliable Data

Determining the accuracy and appropriateness of data rests on the reliability of the data. Reliability is a way to quantify and estimate the consistency of observed scores across multiple observations (Algina & Swaminatahan, 2015; Webb et al., 2006). For example, when an individual's scores vary from test occasion to test occasion or the ranking of the individual derived from the test scores varies among test raters, the dependability of the data along with the defensibility of the use of this data is questionable (Webb et al., 2006). Error is simply a way to describe uncertainty (Vispoel et al., 2018). The concept of reliability gives the assessor a tool upon which to argue the indisputable nature of the collected data. Fundamentally, three questions surround reliability for behavioral assessment: (a) "what affects the reliability of the test?, (b) how can a test be made more reliable?, and (c) what is a satisfactory level of reliability" (Drost, 2011, pp. 105-106)?

There are three main types of reliability: (a) inter-rater reliability, which is the extent to which different raters' judgements are consistent—this might be represented by the Pearson correlation coefficient; (b) test-retest, which is consistency over time and might be represented by the Pearson correlation coefficient also; and (c) internal consistency—the consistency of a person's responses across the items on a multi-item measure, which is typically represented by Cronbach's coefficient alpha, the split-half method, Kuder–Richardson 20 formula, or Hoyt's method (Coulacoglou & Saklofske, 2017; Fan & Sun, 2014; Hauenstein & Embretson, 2019). Of these, Cronbach's coefficient alpha is the most widely used reliability estimate (Fan & Sun, 2014). Generalizability theory acknowledges the same types of reliability but in a much different way. Inter-rater reliability is more about the generalizability of the measurement scores across raters. Test-retest reliability focuses on the generalizability of measurement scores across occasions. Finally, internal consistency can be examined by calculating the generalizability of single facets or the interaction of multiple facets. Moving beyond these types of reliability, other sources impacting reliability include (a) item content heterogeneity, which is a way of assessing the item of the measurement that covers all of the various parts of a trait or focus rather than just a few; (b) retest reliability where retesting influences trait variation over time; (c) item ambiguity, which means the items are difficult to understand and could be a result of arcane and complex vocabulary, ambiguous phrasing, or even contradictory language; and (d) reliability that is impacted by the characteristics of the sample (e.g., teacher bias; Coulacoglou & Saklofske, 2017; Kauffman & Landrum, 2018). All these types of reliability dimensions can be explored through G theory.

Data from behavior rating scales might be abundant with vulnerabilities. With the increased use of behavior rating scales in an applied setting rather than in a controlled laboratory setting comes the need to reassess how error is measured for these instruments (Briesch, Swaminatahan, Welsh, & Chafouleas, 2014). Applied settings are unstructured, uncontrolled, and, therefore, allow error variance in different measurement conditions

(Bruckner, Yoder, & McWilliam, 2006). All variables of the instrument should be evaluated for potential sources of error that might impact the reliability of the scores. Generalizability theory provides a method for this type of scrutiny. Behavior rating scales are one of the most commonly used assessments to identify at-risk students. Moreover, behavior rating scales are used as markers for monitoring social, emotional, and behavioral fluctuations in students over time in addition to providing evaluative data for special education and related services. Understanding variables such as rater type, subscale, item, student age, disability status, and their influence on error impacting the reliability and the validity of the instrument in that it truly measures the construct it intends to measure allows educators to mitigate possible sources of error. Subsequently, this understanding bolsters the accuracy of the data upon which school-based decisions are made and directly impacts the chosen intervention and services chosen to meet the student's needs.

Three common statistical methods for psychometric analysis of reliability and validity are used to define and quantify errors in observed scores: (a) CTT, (b) item response theory, and (c) G theory (Frey, 2018b). Classical test theory, the most commonly applied statistical approach to psychometric testing, does not allow for a greater depth of information regarding sources of error. Item response theory might have advantages over CTT, such as the premise that "response depends on level of ability or skill," but flaws remain as it is a cumbersome analysis and one that is generally difficult to understand (Hays, Brown, Brown, Spritzer, & Crall, 2006; Lord, 2012, p. 12). Before examining generalizability theory more in-depth, the traditionally employed psychometric theory, CTT, is compared to G theory.

Generalizability Theory Compared to Classical Test Theory

According to the *Standards for Educational and Psychological Testing,* all assessments used to gather student data, whether strengths-based or deficit-based, must be psychometrically sound (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational & Psychological Testing, 2014). Chiefly, this involves measures of validity and reliability. Validity is defined as "the interpretation of the observed score as representative of some external property" (Kane, 1982, p. 125). Reliability describes the "consistency of observed scores" (Kane, 1982, p. 125). "The most important concept in classical testing theory (CTT) is the reliability of scores for a given measure" (Vispoel et al., 2018, p. 2). In short, CTT is a means of understanding the relationship between an expected score on a measurement (the true score) and the observed score on a measurement (Hauenstein & Embretson, 2019). Used most frequently for educational research, obtaining stable scores for a particular trait, behavior, or characteristic using CTT is hampered due to its inability to discriminate error sources that could provide critical information to strengthen measurement reliability and validity for diagnostic decisions and progress monitoring purposes (Briesch et al., 2016). Generalizability theory provides a means to alleviate this issue. Table 2 presents a truncated overview of differences and similarities between classical test theory and generalizability theory.

Table 2

Comparisons Among Classical Testing Theory and Generalizability Theory

Issue	Classical Testing Theory	Generalizability Theory
Forms and Parallelism	Classically parallel, essentially tau-equivalent, etc.	Randomly parallel
True Score	Expectation over forms	Expectation over randomly parallel forms
Assumptions	Relatively weak	Very strong
Primary Strengths	Simplicity; widely used; has stood test of time	Conceptual breadth; disentangles multiple sources of error; distinguishes between fixed and random facets
Primary Weaknesses	Undifferentiated error	Conceptual complexity
Use and Understanding	Easy	Sometimes challenging

Adapted from Brennan (2001, p. 19).

Classical Test Theory Constraints

The CTT model is:

$$X=T+e$$

Classical test theory deconstructs the random variable, called score, into three separate categories: (a) observed (*X*), (b) true (*T*), and (c) error (*e*). According to Brennan (2010), CTT has four distinct limitations due in large part to the assumptions derived from the unobserved variables, *T* and *e*. Once it is known what *T* or *e* is, there is an assumption that the opposite *T* or *e* should be evident to the assessor (Brennan, 2010). The true score (*T*) could be described as the expected value of an individual score based on a specific assessment; true score is the average of the observed scores (Salkind, 2010). However, true score does not have a direct connection to the construct the test is intended to measure (Salkind, 2010). Another assumption rests in the emphasis placed on the true score. Replications of the assessment, that is test and retest conditions, result in variations of the true score for the assessment (Brennan, 2010). Furthermore, the variable's true score and error score are unobservable and have no meaning beyond the assumptions assigned to them (Brennan, 2010). Error has to do with extrinsic variables (or the measurement conditions) and their impact on the trait of interest (Hauenstein & Embretson, 2019). These extrinsic variables are potentially rife with error due to human variation (Bloch & Norman, 2012). Generalizability theory, a statistical method used to identify sources of error variance like characteristics of the assessor or test setting conditions, such as time of day, day of the week, season of the year, lighting, or noise, unmasks multiple sources of measurement error (Coulacoglou & Saklofske, 2017; Shavelson et al., 1989).

Among CTT's other constraints as a measurement theory is it does not allow for separate error estimation of potential sources of response inconsistency such as the construct being measured or the occasions of measurement (Hays et al., 2006). Test-retest reliability, inter-rater reliability, and internal consistency reliability measured through the application of CTT assumes the construct or behavior being measured is constant across measurement conditions and that observed variability is due to undifferentiated measurement error (Volpe et al., 2011). For example, the instrument's error score is not able to be understood more deeply due to confounding variables such as rater bias, age, level of education, or socioeconomic status. This creates a limitation impacting feasibility and procedural designs of behavior rating scales and consequently impacts data accuracy and appropriateness in the ongoing quest for valid and reliable data in a data-based decision-making model (Briesch et al., 2016). Generalizability theory breaks free from this barrier to truly understand error variance and its interactions among facets in the g study. To do this, a g study depends on repeated measurements across different conditions (e.g., three raters; Bloch & Norman, 2012).

Score interpretation is influenced by measurement conditions that influence estimates of reliability and the proportion of true score to observed score variance (Shavelson et al., 1989). Random error is derived from extrinsic variables or measurement conditions that vary from replications of the test such as the testing environment, the testing assessor (rater), or the time or the energy of the test participant (Algina & Swaminatahan, 2015; Hauenstein & Embretson, 2019). True score is not independent of other variables and error score is not a residual or model fit error (Brennan, 2010). True score is dependent on the measurement process itself (Hauenstein & Embretson, 2019). Hauenstein and Embretson (2019) described true score as a "purely statistical" measurement of the expected value of the observed score that one can conceptualize as an average of a series of observations (p. 280). Put another way, correlation exists, i.e., the mean deviations in one variable correspond with the mean deviations in another variable (Algina & Swaminatahan, 2015). Finally, Brennan (2010) pointed out that truth and error are not constructs to be uncovered but are very much determined by the rater using these constructs and error does not imply mistake. The rater defines truth and error. Generalizability theory, on the other hand, tries to identify and estimate the degree of the potentially significant sources of error in a measurement.

Error variance is unrelated to true-score variation and treated as random variance (Shavelson et al., 1989).

Traditionally, researchers have conducted assessments measuring unobservable (e.g., trait) as well as observable (e.g., heart rate) phenomena in a controlled, randomized laboratory setting (Briesch et al., 2016). However, the assessment landscape is changing, particularly with regard to an applied setting such as in a community or in a school (Briesch et al., 2014). Federal regulations do not specifically reference by name tiered academic and behavior support structures; nonetheless, schools are using these to provide the full letter of the law. With these structures comes a greater reliance on behavioral data collection via progress monitoring (Briesch et al., 2014). With applied settings such as schools, environmental and assessor variance exponentially grow. The need to differentiate potential sources of error, examine interaction effects among different variables, and expand the rationale for a strengths-based approach to assessment is greater than ever before as the adoption of tiered support structures incorporating progress monitoring and diagnostics continue to be implemented in schools on a larger scale.

An Alternative for Reliability

Generalizability theory, a statistical framework developed by Cronbach, Gleser, Nanda, and Rajaratnam (1972) specifically for the analysis of behavioral assessments, is a broad and flexible framework for examining reliability of a measurement that allows for a greater understanding of error, resulting in the design of complex measurement strategies for the most optimal measurement conditions (Coulacoglou & Saklofske, 2017). The intent of G theory is simple—to examine the dependability of behavioral assessments (Shavelson et al., 1989). Generalizability theory holds advantages over CTT by providing a way to (a) examine multiple sources of variance at the same time in an applied setting by identifying the most egregious sources of inconsistency in response over measurement conditions; (b) inform relative (norm-referenced) and absolute (criterion-referenced) decision making; (c) provide reliability coefficients, also called generalizability coefficients, tailored to the intended usage for the measurement (e.g., diagnostic usage or progress monitoring); (d) develop the most cost-efficient measurement design due to the capacity of G theory to identify major sources of error; (e) increase generalizability of results and understanding of the contexts to which results could be generalized; and (f) provide rationale for the appropriate context, use, and dependability of deficit and/or strengths-based measurement instruments (Briesch et al., 2014; Bruckner et al., 2006; Shavelson et al., 1989; Webb & Shavelson, 2005). Classical test theory asks how accurately observed scores reflect true scores. Rather, G theory asks how accurately these observed scores allow researchers to generalize about the person's behavior in a defined universe of situations (Shavelson et al., 1989). Generalizability to a universe of conditions is of the greatest importance in G theory (Bruckner et al., 2006).

Similar to CTT, assumptions associated with G theory abound. One assumption is random sampling. Facets are randomly sampled from associated universes of admissible observations. What is different from CTT is it seldom occurs in an applied setting. Another assumption is alternate facets could serve the same purpose as other facets because facets are considered random if they are not part of the g study (Vispoel et al., 2018). A third assumption has to do with independence of responses. It is thought a given individual's scores have no bearing on another individual's scores (Vispoel et al., 2018). Furthermore, it is assumed the measurement scale is continuous and measured on an interval scale. Finally, of utmost importance to the design of a study applying G theory analysis is with the recommended completed crossed two random facet design, the number of data points needed is at least 800, which is derived from calculating the number of persons (*p)* x number of items (*i)* x number of raters (*r*; Smith, 1978). However, there is disagreement on this (Webb, Rowley, & Shavelson, 1988). A widely accepted practice for calculating appropriate sample size does not exist (Brennan, 2001; Webb et al., 1988).

Decision studies. Classical test theory views reliability as outcomes from norm-referenced measures "quantifying degrees of consistency of the relative standings of individuals, but not the consistency of actual scores" (Fan & Sun, 2014, p. 14). However, not all decisions are norm-referenced (also referred to as relative decisions). Criterion-reference or absolute decisions denote decisions based on "both the consistency of relative standings of the individuals and the consistency of actual scores" (Fan & Sun, 2014, p. 14). This could be across testing occasions, across two parallel forms, or across two different raters/observers (Briesch et al., 2014). While a Pearson correlation coefficient is used for inter-rater reliability to measure reliability for norm-referenced or criterion-referenced decisions, intraclass correlation coefficients are used (Briesch et al., 2014; Fan & Sun, 2014). Like generalizability theory, intraclass correlation coefficients also rely on the analysis of variance (ANOVA) statistical model to partition score variance into different components for these relative and absolute decisions (Brennan, 2010; Briesch et al., 2014; Fan & Sun, 2014). With intraclass correlation coefficients, a bridge between understanding reliability in CTT and an expanded understanding of reliability as representative of both absolute and relative derived decisions of generalizability theory is revealed.

Universe of admissible observations. In G theory, scores derived from behavioral assessments are “a sample from a universe of admissible observations” consisting of “all possible observations on an object of measurement (typically a person) that a decision maker considers to be an acceptable substitute for the observation in hand” (Webb et al., 2006, pp. 93-94). The heart lies in the accuracy of this generalization from a sample of behaviors to all the possible samples of interest (Volpe et al., 2011). Major improvements over CTT provided by G theory include “estimating main effects and interaction effects for all aspects of the measurement context simultaneously, comparing reliability across combinations of levels of the aspects of a measurement context” (Bruckner et al., 2006, p. 140).

Classical test theory views error as undifferentiated random variation (Webb et al., 2006). In G theory, error variance is divided by its source rather than simply estimated as a single, undifferentiated level of error as is done using CTT (Bloch & Norman, 2012; Tanner et al., 2018; Webb & Shavelson, 2005). This extension of CTT recognizes multiple sources of error as well as overlap or simultaneous error source effects impacting reliability measures (Briesch et al., 2016). Webb et al. (2006) surmised that among the greatest advantages of G theory was the identification, disentanglement, and estimation of sources of measurement error. In many ways, G theory is a comprehensive psychometric measurement theory because reliability methods such as Pearson *r*, coefficient alpha, KR-20, and intraclass correlation coefficients are all represented through this theoretical framework (Fan & Sun, 2014).

Types of measurement errors. Vispoel et al. (2018) described three primary sources of measurement error specific to measurement of individual difference in traits or behaviors: (a) random-response, (b) specific-factor, and (c) transient measurement error. Random-response error could be described as noise affecting scores of a measurement occasion (Bloch & Norman, 2012; Vispoel et al., 2018). Random-response error could be influenced by effort, mood, attention, or memory fluctuations of those being assessed (Vispoel et al., 2018). Specific-factor measurement error is described as when the person being assessed consistently responds to some items unrelated to the construct under investigation (Vispoel et al., 2018). Finally, persistent factors affecting scores during a measurement occasion, such as tiredness or illness but not across occasions, are considered transient measurement errors (Vispoel et al., 2018). When applying G theory, it is important in the analysis of behavioral assessment to consider these potential sources of measurement error seriously as the result could lead to a gross over- or underestimation of reliability.

Consistency, generalizability, and reliability of results. More than one factor (called facet in G theory terminology) could impact the level of generalizability, consistency, and reliability derived from data. Using CTT might lead to reliability issues as it was designed for examining one facet or factor only; it is not able to handle multiple sources of measurement error. With G theory, this is not the case. Generalizability theory is mainly concerned with reliability or “quantifying the consistencies and inconsistencies in observed scores” that “arise or could arise over [multiple] replications of measurement procedure” (Brennan, 2010, pp. 1-2). Reliability through the lens of G theory is “the degree to which a set of measurement values can be repeated under precisely the same measurement conditions, thus reflecting the fundamental question in statistics: “What would happen with the results if I could do the research again?” (Sijtsma & Van der Ark, 2015, p. 128). Providing a theoretical framework and methodology to understand multiple error sources in a measurement, G theory allows for all types of reliability estimates to be calculated such as internal consistency reliability coefficient (Cronbach’s coefficient alpha, KR-20), coefficient of stability (test-retest reliability coefficient), coefficient of equivalence (parallel-form reliability coefficient), interrater reliability coefficient, and intraclass correlation coefficient (estimating measurement reliability across raters and across occasions; Brennan, 1992; Fan & Sun, 2014). In G theory, the generalizability coefficient (*g* coefficient), also known as a reliability coefficient, functions in the same manner as Cohen’s coefficient α and KR-20 with dichotomously scored items (Fan & Sun, 2014). Formulas used to calculate the *g* coefficient are found in Table 3. The *G* coefficient used for relative decision making is conceptually the same as Pearson *r* reliability coefficients, interrater, test-retest, and parallel form (Fan & Sun, 2014). Whereas coefficients derived from G studies that are applied to absolute decisions are conceptually and mathematically the same as intraclass correlations from CTT (Fan & Sun, 2014). This distinction shows the importance for the researcher to clearly understand the type of decision needed for the question(s) at hand. Generalizability theory seeks answers to inter-rater reliability and test-retest reliability differently than in CTT by asking: (a) To what extent do the scores generalize across raters?, (b) To what extent do the scores generalize across occasions?, and (c) To what extent can the scores be generalized across both occasions and raters? Generalizability theory also allows researchers to ask to what extent the scores generalize across items or other assessment items.

Table 3

Formulas Used to Calculate the Generalizability Coefficient

$Xpir =$	Definition
μ	grand mean
$+ \mu p - \mu$	person (student) effect
$+ \mu i- \mu$	item (subscale) effect
$+ \mu r- \mu$	rater (teacher type) effect
$+ \mu pi - \mu p - \mu r - \mu$	person x item effect
$+ \mu r- \mu p - \mu r + \mu$	person x rater effect
$+ \mu ir - \mu i - \mu r - \mu$	item x rater effect
$+ Xpir - \mu pi - \mu pr - \mu ir + \mu p + \mu i + \mu r - \mu$	residual

Traditional aspects of CTT are given a different perspective through the application of G theory: for example, (a) retest or stability where test participants are given the assessment at different times with the same form can be measured through examining the facet of time; (b) equivalence can be viewed through examining the facet, form, such as using two forms to cover the same content with different people; and (c) internal consistency with the facet, item, looking at content heterogeneity and content saturation, or with the facet, rater, to understand fundamental differences between assessors (informants; Coulacoglou & Saklofske, 2017).

Specific to school decision makers, i.e. administrators and school psychologists, applying G theory to behavior rating scales allows individuals to parse procedural inferences impacting rater and item variance for different types of decisions (norm-referenced and criterion-referenced). In turn, this might inform the continued improvement of comprehensive and widely adopted standards for the operational procedures involved in the implementation of universal screens or other behavioral assessments used for school-based decisions.

Examining Reliability and Validity of Behavior Rating Scales

When questions about reliability and/or validity arise for a specific assessment tool, it no longer serves its purpose to inform the allocation of school resources through appropriate and accurate decision-making data (Miller et al., 2017). By examining the variability in student ratings between three rater types using a strengths-based assessment tool (BERS-3) to better understand the factors underlying rater variation might result in improved accuracy of information collected for support services, accommodations in Section 504 plans, and/or individualized education programming. It might also provide more data that directly impact the interventions and services chosen to meet the student's needs. Finally, looking at the BERS-3 from a more global perspective of reliability in terms of norm referenced versus criterion referenced decisions might provide valuable information on how and when this behavior rating scale assessment could be most effectively and efficiently utilized.

Assessment is a means to better understand the deficits of the student and results in data for the purpose of interventions and services chosen to meet the student's unique needs. However, a student is a multi-dimensional being with many strengths in addition to areas for growth. They might show competency in one area yet struggle in another area. Rationale for looking more closely at potential rater variance in a strengths-based rating scale using G theory is a strengths-based approach to assessment said to (a) encourage the student to increase their use of their competencies and strengths; (b) decrease frustrations commonly felt by the student, family, and teachers over the student’s perceived or real inadequacies; (c) increase a student’s overall motivation; (d) bolster new skill development by leveraging the student’s strength areas; (e) support a holistic or comprehensive understanding of the student; (f) extend the focus of the IEP to address preventative interventions and supports; (g) expand the student’s ability to move through adversity and manage stress; (h) foster resiliency; (i) promote positive relationships with others; (j) strengthen self-concept; and (k) celebrate the student’s assets such as a supportive family, an empathetic community, or a student’s positive characteristics (e.g., perseverance or determination; Climie & Henley, 2016; Epstein, 2000; Lambert et al., 2015; Tedeschi & Kilmer, 2005). On the other hand, a deficit-based assessment approach is described as an assessment that scrutinizes a student’s inadequacies “at the expense of recognizing what is going well in individual’s life” (Climie & Henley, 2016, p. 109). The focus is on remediation and providing interventions and supports for behaviors and skills that are lacking rather than on leveraging what the student does well (Climie & Henley, 2016). Concerns regarding a deficit-based approach in diagnostic assessment are it (a) perpetuates already held societal stigmas and stereotypes of students struggling in the social, emotional, behavioral, or mental health domains; (b) cements the professional’s view and focus of the student through a negative lens; (c) encourages the opposite effect of what is desired and might actually disenfranchise and discourage the student; (d) focuses on a limited range of information for consideration in the IEP process; (e) reduces the student’s self-concept; (d) increases the distance between the student and their sense of belonging in the community; and (e) perseverates the student’s past failures (Climie & Henley, 2016; Epstein, 2000; Lambert et al., 2015; Tedeschi & Kilmer, 2005).

Further complicating matters when investigating the dependability of data accrued from behavioral rating scales, whether strengths- or deficit-based, is informants (e.g., teachers or parents) are influenced by attributes, characteristics, and environmental factors (Kaurin et al., 2016). Therefore, they are inherently biased (Millman, 2015). Possible reasons for rater variance leading to biased scoring could be due to (a) age-related limitations or introspection; (b) parental psychopathology; (c) differences in the understanding of which behaviors, states, and traits represent the construct of interest; (d) differences in one’s ability to gain from daily interactions observations that are relevant to the construct being measured; and (e) behaviors, states, and traits not consistent across settings and, therefore, do not allow all raters to observe the same types of behaviors (Kaurin et al., 2016; Millman, 2015). Many studies have been conducted specifically addressing between-rater variance through G theory using behavior rating scales; however, no existing studies specific to behavior rating scales were found that included three-rater types for a strengths-based assessment tool.

Need for the Study

Eighteen studies from 1982 through 2018 were identified as pertinent to this study. All employed the use of generalizability theory for the examination of the reliability of the scores derived from behavioral assessments to improve testing conditions and bolster assessment fidelity whether for universal screening, progress monitoring, or diagnostic purposes. Feasibility was a recurring concern across these studies. Feasibility concerns were attributed to both time constraints and financial burden. None of the 18 studies referenced above used a strengths-based assessment tool in their investigations.

Direct Behavior Rating and Generalizability Theory

Direct behavior rating (DBR) was the instrument used in most of the studies found. Direct behavior rating is "an evaluative rating that is generated at the time and place that behavior occurs by those persons who are naturally present in the context of interest" (Christ, Riley-Tillman, & Chafouleas, 2009, p. 205). Gresham, Dart, and Collins (2017) used an evidence-based behavioral intervention, the Good Behavior Game, which is a direct behavior rating for its instrument with the aim of identifying essential factors to ensure treatment fidelity. Forty-seven children ages one through four who were enrolled in a university-based childcare program participated in a study where the optimal number of observers, sessions, and length of sessions were examined through the lens of G theory to obtain dependable estimates of engagement during free play activities (McWilliam & Ware, 1994). In a similar study, 15 preschool students enrolled in a university-based preschool in rural northeast United States participated in a study examining error variance across raters, time, and setting using DBR (Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese, 2007). In this study, findings indicated DBR was a reliable and valid method for assessing the social behavior of preschool students (Chafouleas et al., 2007). Briesch, Chafouleas, and Riley-Tillman (2010) examined 12 kindergarten students in inclusive classrooms in suburban northeast United States by using DBR and systematic direct observation of single item scales to investigate sources of error across methods, raters, time, and their subsequent interactions. Results from the study showed a minimum amount of direct observation data were necessary to yield reliable low-stakes decisions. A similar study investigating DBR single item scales looking at measures for academic engagement and disruptive behavior was conducted among 7 eighth grade students in an urban northeast United States inclusive classroom (Chafouleas et al., 2010). Error variance across raters, occasions, and days were of most interest. Findings echoed the results from Briesch et al. (2010) that reliability measures derived from DBR could be dissected and untangled to bolster generalizability. In another study examining 9 seventh grade students using DBR in an urban, public charter school in northeast United States, Volpe and Briesch (2017) were able to isolate the occasion facet to decrease feasibility concerns and formulate appropriate implementation procedures for the DBR.

Examining an elementary school population of 14 fifth grade students in an inclusive classroom in northeastern United States, Hintze and Matthews (2004) investigated generalizability theory applied to direct observations of student behavior across time and setting specifically examining student engagement. Continuing this study, Volpe, McConaughy, and Hintze (2009) used a direct observation form to examine on-task and off-task behaviors applied to a narrower population—students who were all referred for learning and behavioral and emotional functioning deficits. In this study, 24 six-year-old through 11-year-old students across 18 elementary schools in New England were included. Time and setting were variables of interest with research questions related to how much time was necessary for reliable results and which types of classroom settings resulted in higher levels of reliability on the direct observation form. A few studies specifically retained undergraduates or researchers to carry out the direct behavior observation using video clips of students that were analyzed later (Christ, Riley-Tillman, Chafouleas, & Boice, 2010; Wickard & Hulac, 2017). Ratings from individual or small groups of simultaneous raters, when generalized only to that specific individual or group of individuals, met reliability criteria for both low and high stakes decisions (Christ et al., 2010). Christ et al. (2010) found that decisions, both norm-referenced (relative) and criterion-referenced (absolute) or in more common language for universal screening, progress monitoring, or diagnostic purposes, could achieve levels of acceptable reliability when manipulating rater variables.

Between and Within-Rater Variance

Rater variance and, more specifically, teacher rater variance was investigated thoroughly in the located studies. Millman (2015) found a 20.5% variance in between-teacher at-risk assessment scores citing that most office discipline referrals originating from the same group of teachers, teacher reports, whether academic or behavior based, were not necessarily independent from the actual teacher completing the assessment. Likewise, Smith-Millman et al. (2017) posited one single teacher typically completed the assessment for that one student and factors like classroom differences or rater biases might produce varying scores. Gage, Prykanowski, and Hirn (2014) studied intra-observer (rather than inter-observer) agreement. They investigated within-rater agreement across time and occasions using direct behavior observation ratings of seven teachers working with kindergarten through fifth grade students from three elementary schools in an urban, southeast U.S. school district. A central aim was to mitigate low reliability that might result from within-rater biases' over time, occasions, and their interactions because these were associated with an increased probability for a Type II error (Gage et al., 2014). A Type II error is a "false negative": failing to reject a false null hypothesis. Following suit in a study conducted in seven elementary schools with 68 mainly female teachers who were overwhelmingly Caucasian and 1,241 mostly African American students with equal gender representation, evidence showed significant between-teacher variance in ratings of student behavior predicted by student, teacher, and classroom demographic characteristics; the teacher's professional development; and academic performance characteristics of the class (Splett et al., 2018). Furthermore, results from a study of 1,700 student participants from first, second, fourth, fifth, seventh, and eighth grades examining mean score performance using Direct Behavior Rating-Single Item Scales indicated 20% to 24.6% of variance was due to teacher and classroom level differences (Johnson et al., 2016). Mean-group differences emerged from a study conducted by Peters et al. (2014) with a sample of 982 students in kindergarten through fifth grade whose teachers completed the Clinical Assessment of Behavior-Teacher Form (Bracken & Keith, 2004). Significant variance was due to teacher- and school-level variables. Teacher self-efficacy for behavioral management along with teacher age were two facets contributing to this error variance. Previous studies have focused on between-teacher variance on behavioral assessments administered to elementary–age students (Mason et al., 2014; Rowe et al., 2019; Smith-Millman et al., 2017; Splett et al., 2018; Tanner et al., 2018; Wolcott & Williford, 2018). No studies have been conducted thus far to examine between-rater variance for middle and high school students.

Item and Occasion Sources of Error Variance

Used as formative assessments like progress monitoring tools, behavior rating scales have been criticized as having too many items with too many directions along with too great of a financial burden (Chafouleas et al., 2007, 2010). This directly impacts adoption of the assessment in the school setting (Chafouleas et al., 2007, 2010). Additionally, if these behavior rating scales have been shortened or adapted for formative assessment, questions arise regarding how thoroughly these new versions of the assessment have been investigated and tested to ensure the items or tasks within the assessment reflect appropriate levels of reliably and validly of the construct of interest (Chafouleas et al., 2007). Volpe et al. (2011) conducted a study examining the IOWA Conners Teacher Rating Scale (Pelham, Milich, Murphy, & Murphy, 1989) that included 71 participants ages 6 to 13-years-old. Volpe et al. was particularly concerned with how frequently raters were asked to provide ratings and how the number of items or the length of the rating scale impacted rater load when used for progress monitoring. One-third of the observed rating variance was explained by the relative standing of students across the conditions like person by condition and person by occasion with condition facets indicating undesirable variation (Volpe et al., 2011). This negatively impacted the generalizability of the resulting scores. Of interest was this study examined results of G theory through the lens of both the relative decision study outcomes and the absolute decision study outcomes. As mentioned previously, this was particularly important when looking at the reliability of the scores derived from a measure for multiple uses such as progress monitoring and diagnostic purposes.

Instrument Variance

Bergeron, Floyd, McCormack, and Farmer (2008) investigated externalizing behavior error variance across occasion, rater, instrument, and interaction through the use of the Behavior Assessment System for Children Second Edition (Reynolds & Kamphaus, 1992), which are brief, targeted forms and software for monitoring changes in behavior or emotional status, and the Achenbach System of Empirically Based Assessment (Achenbach, 2014), a comprehensive assessment system for both formative and summative behavioral assessments. Of interest in this study was the examination of specific source variance between instruments. Instrument variance "refers to inconsistencies between scores that supposedly assess the same construct (e.g., aggression) yielded by different rating scales administered concurrently" (Bergeron et al., 2008, p. 92). Response format and item wording differences between two instruments measuring the same constructs might have led to this variance. Bergeron et al. (2008) acknowledged that even though there was not a widely accepted level of acceptable consistency between scores from multiple rating scales administered at the same time, the dependability of the scores derived from the instruments should be moderate or higher in magnitude. Less than 3% total variance for externalizing behavior composites was found between the Behavior Assessment System for Children Second Edition and the Achenbach System of Empirically Based Assessment. This variance grew to between 8% and 17% of the total variance when looking at subscales of the instruments. An explanation given for the larger subscale error variance was the number of items for the subscales from both instruments varied widely from five items to 20 items. Alternately, if looking at externalizing behavior items only, both instruments had the same number of items representing this construct—30. The large subscale variance was attributed to a similar phenomenon in classical test theory, the Spearman-Brown prophecy formula, which indicated a longer test had an increased internal consistency; i.e., the "aggregation of ratings from a greater number of items" result in "stronger dependability coefficient for the externalizing composites" (Bergeron et al., 2008, p. 102). An alternate rationale for the findings was differences in teacher judgements due to different language used to describe the items measuring the same constructs for the two assessments.

Behavior Rating Scales

Five studies were located examining generalizability theory and behavioral assessment using behavior rating scales as the instrument of interest (Bergeron et al., 2008; Conger, Conger, Wallander, Ward, & Dygdon, 1983; Crowley, Thompson, & Worchel, 1994; Rowe et al., 2019; Tanner et al., 2018). Two of the three studies employed instruments for diagnostic and identification purposes that included only elementary-aged students (Bergeron et al.; Conger et al., 1983) and one included students age 11 through 16 years in their study of G theory applied to diagnostic assessment (Crowley et al., 1994). Rowe et al. (2019) conducted a large study with 1,100 participants in kindergarten through fifth grade (52% girls and 48% boys), 50% of whom were African American in four different elementary schools over a three-year period. This was the only study located investigating behavioral assessment and G theory conducted over multiple years. Student x teacher x occasion was the design of the study, which intended to look more fully into potential rater variance impacting reliability estimates. Rowe et al. were particularly concerned with relative, low stakes decision making such as progress monitoring. Conger et al. (1983) used the Conners Teacher Rating Scale (Conners, 1969) as the instrument to examine rater variance according to rater type, which was an outside assessor or classroom teacher. Finally, Crowley et al. (1994) looked at the g study design, person x item x occasion, for sources of error variance using the Children’s Depression Inventory (Kovacs & Beck, 1977) among 164 children in Texas. A follow up decision study (d study) was not conducted; therefore, the findings were limited as a d study is crucial to the application of G theory. Finally, investigating differences among students, raters, occasions, and screening measures impacting reliability data derived from universal screening procedures was of concern to Tanner et al. (2018). For this study, three female teacher-pairs from a suburban middle school in southwestern United States rated 82 students in sixth, seventh, and eighth grades who were mainly Caucasian/White using the Social, Academic, and Emotional Behavior Risk Screener-Teacher Rating Scale (Kilgus & von der Embse, 2014), a universal screen used for assessing student risk for social, emotional, and behavioral problems among students in kindergarten through 12th grade, and the Strengths and Difficulties Questionnaire-Teacher Form (Goodman, 1997), a mental health screening tool for children and adolescents 2 years old through 17 years old used by researchers, clinicians, and educators. None of the studies located included the BERS-3 (Epstein & Pierce, in press) and no measurements used were from a strengths-based assessment approach. These instruments were of interest in this study specifically to examine and understand more completely rater variance and subsequent progress monitoring and potential diagnostic decision-making dependability when using a strengths-based assessment approach. Incorporation of the BERS-3 as the instrument in this study was integral to exploring more fully the reliability and validity of the outcomes of the scores derived from this assessment approach. Understanding reliability concerns specifically examining rater variance might directly impact the special education and related services chosen to serve the needs of students with ED.

Summary

Previous studies have focused on between-teacher variance on behavior rating scale scores administered mainly to elementary–age students (Mason et al., 2014; Rowe et al., 2019; Smith-Millman et al., 2017; Splett et al., 2018; Tanner et al., 2018; Wolcott & Williford, 2018). Additionally, a large percentage of studies that investigated reliability of behavior assessment tools through generalizability theory—whether looking at rater, item/task, or occasion facets for reliability levels appropriate for norm-referenced or criterion-referenced decisions—focused on preschool children only (Chafouleas et al., 2007; McWilliam & Ware, 1994), elementary-aged students only (Hintze & Matthews, 2004; Lomax, 1982), and middle school students only (Chafouleas et al., 2010; Volpe & Briesch, 2016). A few studies incorporated both elementary and middle school students in one study (Bergeron et al., 2008; Crowley et al., 1994; Volpe et al., 2011). No studies have been conducted thus far to examine rater variance using a strengths-based behavior assessment. Furthermore, of the studies available investigating G theory and behavior rating scales, no studies included students in both middle and high school in one study. Likewise, no generalizability studies were found that included a population of high school students beyond the ninth grade. To contribute to the body of knowledge, in this dissertation, rater-level variance on a strengths-based behavior rating scale administered to middle and high school students and used for progress monitoring purposes was examined through G theory. The scores obtained from (a) the students rating themselves, (b) the middle or high school teacher working directly with the student, and (c) one of the students' parents were compared to investigate rater-agreement.

CHAPTER III

METHODOLOGY

Introduction

This chapter outlines the methods used, procedures, data analyses conducted, as well as the rationale for selecting this approach to answer the following research questions.

Q1 To what extent does rater type (e.g., student, parent, and teacher) explain rater scores derived from a strengths-based behavioral rating scale?

Q2 To what extent are the scores derived from three rater informants on a strengths-based behavioral rating scale reliable for use in absolute or relative decisions?

It was hypothesized there would be differences in scores from the BERS-3 (Epstein & Pierce, in press) by rater type (student, parent, and teacher) that significantly impacted access to educational supports and services for students with emotional disturbance.

Participants and Data Collection

I contacted PRO-ED (2020) via email and shared with a research analyst an overview of the purposed study to see if the authors of the BERS-3 (Epstein & Pierce, in press) and PRO-ED would be willing to grant permission for the use of the norming data they were collecting for the third edition of this instrument. Permission was granted Summer of 2019 (see Appendix A for proposal sent to PRO-ED for permission to access norming data for third edition of the BERS rating scale and Appendix B with communication to and from PRO-ED granting permission for access to the data). PRO-

ED provided copious data sets. Of the initial 228 cases, 25 were found to meet the criteria for this study using the BERS-3 matched to three rater types for students in middle and high schools.

The BERS-3 (Epstein & Pierce, in press) normative data were collected from Fall 2015 through Spring 2018 and were a weighted sample representing 1,430 children ages 5 years 0 months through 18 years 11 months and from 23 states and 180 different zip codes. Represented states were Arizona, California, Colorado, Florida, Georgia, Iowa, Kansas, Louisiana, Maine, Maryland, Massachusetts, Montana, Nebraska, Nevada, New Hampshire, New Jersey, New York, North Carolina, Texas, Utah, Vermont, Washington, and Wyoming. With regard to geographic region, gender, race, Hispanic status, exceptionality status, parent education level, and household income, the sample of student evaluated was representative of the nation as a whole. Percentages for these characteristics were compared with those reported in the *ProQuest Statistical Abstract of the United States, 2016* (ProQuest, 2016) and the *Digest of Education Statistics 2015* (Snyder & Dillow, 2015). From the complete data set provided from PRO-ED, only cases using the BERS-3 that had one student assessed by three rater types were included in this study. Table 4 provides specific demographic characteristics of this smaller, more focused data set.

Table 4

Demographics of the Three Rater Forms

	n	Percentage of Study Sample	Percentage of U.S. School-Aged Population[a]
Age (in years)			
11	4	16	-
12	1	4	-
13	2	8	-
14	6	24	-
15	5	20	-
16	3	12	-
17	3	12	-
18	1	4	-
Geographic Region			
Northeast	0	0	16
Midwest	6	24	21
South	12	48	38
West	7	28	24
Gender			
Male	19	76	51
Female	6	24	49
Ethnicity			
White	16	64	73
Black/African American	6	24	15
Asian/Pacific Islander	0	0	5
American Indian/ Alaska Native	0	0	2
Two or more	3	12	5
Hispanic Status			
Yes	1	4	24
No	24	96	76
Exceptionality Status			
Intellectual disability	1	4	1
Attention deficit/hyperactivity disorder	5	20	10
Articulation disorder	1	4	3
Language impairment	3	12	3
Emotional disorder	2	8	1
Behavior disorder	10	40	1
Low functioning autism	2	25	1
High functioning autism	1	4	1
Developmental delay	1	4	1

[a]Based on data reported in the ProQuest *Statistical Abstract of the United States, 2016* (4th ed.), by ProQuest LLC and Bernan Press, 2016, Bethesda: Author.

Items omitted by the rater or items with more than one response posed a problem for scoring. The examiner (e.g., those conducting the collection of the norming data) encouraged raters to answer all items and reminded them to mark one response only. If too many items were omitted or multi-marked items occurred, the validity of the rating scales would be undermined. On the BERS-3 (Epstein & Pierce, in press) teacher, parent, or student forms, two omitted or multi-marked items were allowed per subscale for the five core subscales; more than two rendered that specific subscale unscorable. Scoring for the other subscales still occurred.

Any omitted or multi-marked items on the supplemental Career Strength subscale canceled the use of the scale. Scores for the unscorable items were estimated by computing the average of the other items in the same subscale. For example, if two items were missing for the Interpersonal Strength subscale, the average (rounded to the nearest integer) of the remaining 13 items that made up that scale was computed and added twice (once for each unscorable item) in computing the raw score for that subscale. A total of 10 unscorable items (two for each of the five core subscales) was permitted for the BERS-3 (Epstein & Pierce, in press).

Instrumentation

For this study, the dependent variables (or facets in G theory) were the scores derived from each rater on each subscale and strengths index score signaling the student may have ED and needed follow up data gathering and reporting or the student needed intervention, supports, special education or related services for social, emotional, and behavioral areas currently impacting their access to education. The independent variables were rater facets for student, teacher, and parent.

The rationale for obtaining secondary data for the BERS-3 (Epstein & Pierce, in press) began with the graduate work I participated in that centered on the use of this assessment tool. Duties included data collection, data entry, and collaboration on written introductions for two manuscripts with Dr. Corey Pierce, my advisor. A pilot study evaluating variance between-and within-different teacher types (e.g., general education teacher and special education teacher) through the lens of generalizability theory was conducted in Spring 2019 using the Strengths and Difficulties Questionnaire (Goodman, 1997). This opportunity allowed me to gain more knowledge about G theory, to experiment using different types of G theory statistical software packages, and to understand more about score reliability and its impact on a data-based decision-making process. The BERS-3 rating scale served a requirement for my study to include a strengths-based (BERS-3) behavior rating scale in this investigation (see Table 4 for demographic information related to the matched samples on the BERS-3 used for this study).

The Behavioral and Emotional Rating Scale-3 (Epstein & Pierce, in press) is a standardized, norm-referenced assessment of behavioral and emotional strengths for adolescents ages 5 through 18. Table 5 provides instrument information. Table 6 provides information for interpretation of standard scores and Strength Index. Of importance to this study is this assessment investigates behaviors and emotions through a strengths-based approach rather than from a deficit-based approach.

Table 5

Behavioral and Emotional Rating Scale-3: Scales and Scores

PRS, TRS, YRS Subscales	Number of Items	Likert Rating Scale
Interpersonal Strength	15	0 = Not at all like
Family Involvement	10	1 = Not much like
Intrapersonal Strength	11	2 = Like
School Functioning	9	3 = Very much like
Affective Strength	7	
BERS-3 Strength Index	52	

Note: There are 52 total statements contained within the BERS-3. Each statement associated with the subscale are added together to result in one composite score, Strength Index. PRS = Parent Rating Scale, TRS = Teacher Rating Scale, YRS = Youth Rating Scale.

Table 6

Interpretation of Subscale Standard Scores and Strength Index

Behavioral and Emotional Strength	Subscale Subscaled Scores	BERS-3 Strength Index	Probability Student has ED
Very Superior	17-20	>130	Extremely Low
Superior	15-16	121-130	Extremely Low
Above Average	13-14	111-120	Very Low
Average	8-12	90-110	Low
Below Average	6-7	80-89	High
Poor	4-5	70-79	Very High

Note: There are 52 total statements contained within the BERS -3. Each statement associated with the subscale are added together to result in one composite score, Strength Index.

The normative sample for the BERS-3 (Epstein & Pierce, in press) was a weighted sample composed of 1,430 adolescents ages 5 years 0 months through 18 years 11 months, who lived in 23 states and 180 different zip codes. The BERS-3 is composed of three rating scales: (a) Parent Rating Scale (PRS), (b) Youth Rating Scale (YRS), and (c) Teacher Rating Scale (TRS). The PRS and TRS contained 52 items rated on a 4-point Likert scale that ranged from 0 (*Not at all like the child*) to 3 (*Very much like the child*) and was designed to be completed in about 10 minutes (Epstein, 2004). All three types of BERS-3 rating scales were written at a fifth-grade reading level to ease application purposes (Buckley & Epstein, 2004). The rater read each statement and marked the rating that reflected the characteristics representative in the child being rated. The scales were composed of an overall Strength Index, which is a summary score of the (a) Interpersonal Strength, (b) Family Involvement, (c) Intrapersonal Strength, (d) School Functioning, and (e) Affective Strength subscales. A mean standard score of 10 and a standard deviation of 3 applied to all five subscales. The sum of the subscale standard scores was converted into the Strength Index that had a mean of 100 and a standard deviation of 15. In addition to the Likert-scale questions, the scale also contained eight open-ended questions that allowed parents and teachers to note a child's specific academic, social, athletic, family, and community strengths. The PRS included one additional subscale, Career Strengths, that measured the career and vocational strengths of the rated child. The YRS was completed by adolescents ages 11 through 18 and was identical to the parent and teacher scales except for word changes to reflect a student's perspective. For example, "asks for help" was changed to "I ask for help when I need it."

The five-item Career Strengths subscale in the PRS was included in the YRS as well. The 57 items in the YRS could be typically completed in about 10 minutes.

Behavioral and Emotional Rating Scale-3: Reliability

Reliability coefficients for the behavior rating scales had to reach or exceed .80 in magnitude to be considered reliable. Reliability coefficients above .90 were more desirable. Four types of reliability coefficients—coefficient alpha, test–retest, interrater, and scorer difference—were calculated to examined reliability of the scores derived from the BERS-3 (Epstein & Pierce, in press). Content sampling was investigated through the application of Cronbach's (1951) alpha method. Coefficient alphas for the TRS and PRS subscales and composite were calculated at 14 age intervals using data from the entire normative sample. Coefficient alphas for the YRS were calculated at eight age intervals using data from the entire normative sample. The average coefficients for the TRS subscales ranged from .89 to .96 (median = .93). The average coefficient for the TRS composite was .98. The average coefficient for the PRS composite was .97. The average coefficients for the YRS subscales ranged from .80 to .88 (median = .82). The average coefficient for the PRS composite was .94. The SEM estimated the amount of error in a student's score due to less-than-perfect reliability of a rating scale. The SEM was based on the formula (*SD* is the standard deviation for the score of interest [3 for subscales, 15 for composite]). The SEM for the subscales was 1 for the TRS, PRS, and YRS. The SEM for the composite was 2 for the TRS, 3 for the PRS, and 4 for the YRS. The smaller the SEM, the more confidence one could have in the rating scale's results. Performance by sample with emotional and behavioral disorders and matched sample for the TRS ad YRS indicated a moderate to large magnitude and a large magnitude for the PRS.

Behavioral and Emotional Rating Scale-3: Validity

Three types of validity studies were conducted on the BERS-2 (Buckley & Epstein, 2004) norming data: (a) content validity, (b) criterion-related validity, and (c) construct validity. Discrimination indices recommended by Ebel (1972) and Pyrczak (1973) were followed and resulted in item discrimination coefficients exceeding the .35 value for all subscales. The mean coefficients showing the relationship between the BERS-2 subscales and composite scores with the criterion test were in the moderate to large range. Finally, content validity was measured using factor analysis. Results indicated the five-factor model was well established and fit the data well: an adjusted goodness-of-fit index of .965, a normed fit index of .961, and a relative fit index of .959.

Behavioral and Emotional Rating Scale-3: Item Response Theory Measures

Through the lens of item response theory, a differential item factor analysis was conducted for each of the rater types included in the BERS-3 (Epstein & Pierce, in press). The TRS indicated when examining male versus female students there were three statistically significant items on the Interpersonal subscale, four statistically significant items on the School Functioning subscale, and one statistically significant item on the Affective Strength subscale. Statistical significance was found when examining Black or African American students versus non-Black or non-African American students on one item of the School Functioning subscale. Finally, for the TRS, statistical significance was found when examining Hispanic versus non-Hispanic students on one item of the Family Involvement subscale, two items on the Intrapersonal subscale, and two items on the School Functioning subscale. The (PRS was found to have one item each on the Interpersonal subscale, Family Involvement subscale, the Intrapersonal subscale, and the Affective Strength subscale when controlling for Black or African American versus non-Black or non-African American students. Statistical significance was found on one item on the Interpersonal subscale and the Intrapersonal subscale and two items on the School Functioning subscale when examining male versus female students. Only one item indicated statistical significance on the Family Involvement subscale when investigating Hispanic versus non-Hispanic students. Finally, using the YRS, statistical significance was found when examining male versus female students on one item of the Interpersonal subscale, School Functioning subscale, and Affective Strength subscale and two items on the Intrapersonal subscale. Using the YRS, statistical significance was found on two items of the Interpersonal subscale when controlling for Black or African American versus non-Black or non-African American students. Lastly, one item on the YRS Family Involvement subscale was statistically significance when controlling for Hispanic versus non-Hispanic students.

Procedure

Before analyzing the secondary data from PRO-ED (2020), appropriate approval from the Institutional Review Board (IRB) of the University of Northern Colorado was received (see Appendix C for IRB approval letter). Prior to IRB approval, PRO-ED was contacted to inquire how to retrieve the data. PRO-ED removed all identifying information from all data spreadsheets and documents shared; provided a series of tables with demographic information, validity data, and reliability study related outcomes; and the unpublished third edition manuals, second edition manuals, and copies of assessment forms for the BERS-2 and BERS-3 rating scales were either mailed to my home address or sent in an electronic format via email. An Excel datasheet was sent electronically to me with the raw scores for all items listed separately. In addition to raw data from teacher forms from each assessment matched to one student, the BERS-3 (Epstein & Pierce, in press) data contained information on gender, race, ethnicity, date of birth, zip code, and disability status.

Secondary Data Procedures

Educators from four U.S. regions—Northeast, Midwest, South, and West—participated in the norming study for the PRS, YRS, and TRS. For the purposes of this study, they were referred to as Parent, Student, and Teacher, respectively. Authors of these scales contacted participants via email or telephone. Participants signed an agreement to complete the scale on all of their students or to select an unbiased sample of their students. The scale authors gave the following instructions to participants to bolster the probability for an unbiased sample of students:

> First, decide how many students you wish to rate. Then, start at the top or bottom of your class roster and rate every child. Do not skip any child unless you have known this child for less than two months. Stop selecting and rating children when you have reached the number of children you wished to rate (Epstein et al., in press).

Additionally, the scale authors used two data quality control procedures during this process:

1. Trained staff at PRO-ED inspected completed rating scales. Scores and subset data were scrutinized to ensure adherence to scoring rules.

Incomplete demographic information or missing rating scale items excluded the scale from the study.

2. Statistical programming was used to check data electronically for potential irregularities. Discrepancies were compared between the electronic and the paper version as appropriate.

Generalizability Theory Procedures

The following procedural recommendations for generalizability theory-based studies from Briesch et al. (2014, 2016) were followed in this study:

- Described the measurement procedure including when data were collected and by how many people,
- provided component variance clearly allowing the reader to identify the sources of variance contributing to the greatest percentage of measurement error,
- identified the universe of generalization, presented summary statistics, described the object of measurement and raters, and reported estimation procedures,
- specified whether the facets in the study were random or fixed. The purpose for this was the nature of the facet directly impacted generalizability beyond the current studies’ components of scale used and raters,
- included descriptive data for all facets,
- described participants in the study ensuring that participants were representative of those whom the study was intended to generalize,

- completed the scale using raters who were the intended population of users. Researchers should not complete the scales,
- clarified and described conditions where the results might generalize in both the methods and discussion sections of the study,
- described data analysis, assumptions, and types of statistical software employed to investigate the data to ensure appropriateness of evaluation and replicability, and
- explored a wide array of d study manipulations to provide a multitude of measurement and generalizability scenarios for the intended users.

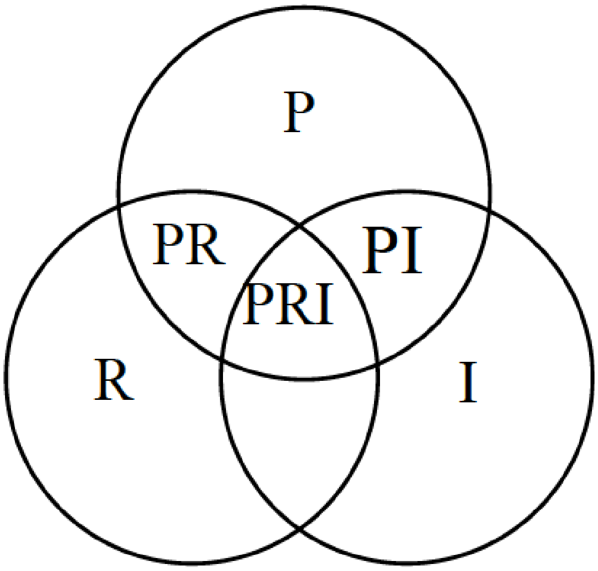
Data Analysis

Often, classical test theory is employed to analyze the interrater reliability of scores obtained from psychometric tests and assessments (Traub, 1997). However, in this study, I analyzed test variance through generalizability theory (G theory), which presumed that error was systematic and could be multi-faceted. This type of error allowed investigation of variance that might be due to procedural issues regarding how the assessment data were collected (Tanner et al., 2018). The end goal was to gather information on the dependability of strengths-based behavioral assessments for potential diagnostic and subsequent progress monitoring purposes used directly for special education and related service decision making for students with ED.

Study Design

To design this study, several components were essential to determine after identification of the dependent and independent variables. The basic elements included (a) the object of measurement or facet of differentiation—the student, (b) facets of generalization—the rater type and the subsequent scores derived from the rater type, (c) stratification facets—none of which were identified, (d) nature of the facets—they were crossed in this study, and (e) number of levels for each facet—three levels of rater type and six levels of subscales, which included the overall Strength Index (Bloch & Norman, 2012). After these basic elements were decided upon, the problem under investigation was determined, the data organized, and the g study analysis for group means, mean-square difference for groups, group variance estimates, variance components for effects estimates, and appropriate *g* coefficients were calculated along with descriptive statistics and ANOVA (Bloch & Norman, 2012). The ANOVA statistical model was the basis upon which the total score variance was divided into variance components of different sources (Fan & Sun, 2014). For example, in this research study, the design involved an object of measurement (*p* or person) and two facets: rater type (*r*) and subscale (*i*). Facets are the same as factors in an ANOVA model and can be fixed or random (Vispoel et al., 2018). This design called for the use of a three-way ANOVA model where there were three main effects (*p*, *r*, *i*), three two-way interactions (*p* x *r*, *p* x *i*, *r* x *i*), and one three-way interaction (*p* x *r* x *i*) confounded with error (*e*). The percentage of variance attributable to each facet aided in the understanding of the magnitude of the measurement error from different sources.

Both SPSS v.25 and EduG v.6 were used for this analysis. One important note regarding the determination of an appropriate data set size is necessary. Webb et al. (1988) acknowledged a widely accepted practice for calculating appropriate sample size does not exist. However, they recommended a minimum of 20 persons and two conditions per facet totaling 40 data points (Webb et al., 1988). This crossed-research g study design was represented as person (students) x rater type (student, teacher, and parent) x item or subscale or *p* x *i* x *r* (see Figure 1 for partitioning of variance and study design). Total data points for this study were determined by calculating the smaller data set extracted from the main secondary data set provided by PRO-ED (2020): 25 for *p* x 3 for *r* x 6 categories for *i*. Total data points were 450.



P = Object of measurement: Student. Proportion of variance in individuals' observed scores.
R = Rater Type, representing scores dependent upon the person measuring the student.
I = Item or Subscale representing the five subscales and one overall Index scaled scores of the BERS – 3.

Figure 1. Venn diagram showing the partitioning of variance for this two facet fully crossed generalizability theory study (adapted from Brennan, 1992).

For the norming data PRO-ED shared with me, the BERS-3 (Epstein & Pierce, in press) had one teacher and one parent score the same student. That student then scored themselves (Youth Rating Scale). Of interest was any systematic variability in subscale and over strengths index according to type of rater using a strengths-based behavior rating scale (BERS-3) followed by a comparison of variance across all rated students. Magnitude of variability was estimated and compared to make decisions about the adequacy of the measurement. This provided evidence regarding the extent of generalizability as well as any variance due to the object of measurement and several potential sources of error. This study used mean squares and *N* values from an AVOVA conducted on the fully crossed scores to estimate the reliability of the obtained between-rater and item variance on the subscales or construct of interest. Table 7 lists equations used for variance components.

Table 7

Equations for the Variance Components of the Generalizability Study

Source of Variation	Variance Component	Equation for calculation from mean squares (MS)
Persons (p)	σ^2p	MSp – MSpi – MSpr + MSpri, *e*/(nr*ni)
Person x rater (pr)	σ^2pr	
Person x item (pi)	σ^2ps	
Person x rater x item, error (pri, *e*)	σ^2pri, *e*	

Note: Adapted from Shavelson and Webb (1991); MSp = mean square person; MSpi = mean square person x mean square item; MSpir, *e* = mean square person x item x rater + error; nr = number of raters (teacher types); ni =number of items.

Researchers should employ a fully crossed design with many representative levels of each facet rather than a partially nested or completely nested design (Brennan, 2001; Shavelson et al., 1989). A fully crossed design is one in which selected levels reflect the characteristics of the relevant universe of the facet to obtain the highest levels of d study accuracy (Brennan, 2001). It could be described as a design in which there is a score for every person (subject) on every level of every facet (item, instrument, occasion, etc.; Bruckner et al., 2006). A fully crossed design was used for this study.

Mean squares and *N* values from an ANOVA performed on the fully crossed scores obtained from between-subject variances on the continuous measure of interest were used to estimate reliability (Bruckner et al., 2006). Participants were the object of measurement; therefore, the variance component for participants was anticipated to be large because people differed in their aptitude/ability on the variable of interest (Bruckner et al., 2006). Each dependent variable had a separate ANOVA. A three-way ANOVA was conducted with items and raters as the facets and people as the objects of measurement. The variance component option using SPSS v.25 to conduct ANOVA gave an estimated variance of the object of measurements, facets of the measurement, context, and the two- and three-way interactions.

Generalizability and Decision Study Coefficients

Generalizability and *d* coefficients were also calculated for this study using EduG v. 6 statistical software. Generalizability and *d* coefficients are similar in nature to the reliability coefficient in CTT, which is the "ratio of the universe-score variance to the expected observed-score variance" (Webb & Shavelson, 2005, p. 605). For this study, raters and items were the facets of the design and comprised the average of scores across rater type and subscales within the students as represented in the following formula:

$$\frac{\sigma^2 p}{\sigma^2 p + \sigma^2 pr + \sigma^2 pi + \sigma^2 pri, e}$$

The universal score variance of person (student) is represented by σ^2p, σ^2pr represents universal score variance of person x rater, σ^2pi represents universal score variance of person x item (subscale), and σ^2pri, *e* represents the universal score variance of person x rater x item confounded with unsystematic or unmeasured error (Morgan, 2001; Webb et al., 2006). The *g* coefficient for relative decisions and *d* coefficient, or index of dependability, for absolute decisions were calculated through EduG v.6 statistical software using the formulas, respectively:

$$p^{\mathbf{2}} = \begin{matrix} \sigma^2(\tau) \\ \sigma^2(\tau) + \sigma^2(\delta) \end{matrix}$$

$$\phi = \begin{matrix} \sigma^2(\tau) \\ \sigma^2(\tau) + \sigma^2(\Delta) \end{matrix}$$

The mean square for person (e.g., object of measurement), the mean square for the person x item interaction, the mean square for the person x rater interaction, the mean square for the person x rater x item, across rater variance, the number of participants in the g study, the number of fully crossed raters, and the number of fully crossed items were inputted into EduG v.6. Of note, normal distribution across facets for the variance components derived from ANOVA is assumed in g studies (Bruckner et al., 2006). However, tests for normal distribution of the data, the Kolmogorov-Smirnov and Shapiro-Wilk, were analyzed to understand more fully the extent of this assumption.

The generalizability coefficient's and the dependability coefficient's magnitude were evaluated to assess the relative and absolute variances (Webb et al., 2006). The sum of adjusted error variances in relation to the person (object of the measurement) was the relative variance (Naumenko, 2015). Absolute variance was the sum of all the variances crossed with or because of the object of the measurement (Naumenko, 2015). A universal estimate of the dependability of the data that considered any measurement variance in the design of the g study resulted from these variance components (Morgan, 2001; Naumenko, 2015). For the purposes of this study, reliability measures for both relative decisions and absolute decisions were conducted as both progress monitoring decisions and potential diagnostic decisions were under investigation.

Descriptive Statistics

Descriptive statistics were performed using SPSS v.25. Although descriptive statistics were not necessary for this g study analysis (Brennan, 2001; Shavelson et al., 1989), these were computed so I would have a better understand of the data and a clear picture of the data through the traditional lens of CTT. Mean differences between rater type and subscale scores for each student were calculated as well as standard deviations (*SD*), range, standard error of measurement (*SEM*), and variance. SPSS v. 25 was used to calculate eta squared values, test for homogeneity, and normality of data.

Generalizability theory analyses might be descriptive only (Briesch et al., 2014). Statistical significance was not relevant (Briesch et al., 2014). Consequently, no statistical tests were conducted for this study for the purpose of analyzing the data through the lens of G theory. I focused on estimating overall *g* coefficients as well as variance components and the percentage each contributed to the score variability using

EduG v.6. If the scores derived from the different types of raters were reliable, the person variance component should represent the largest percentage of variance out of all the sources of variance given that students were the object of measurement. Variance from other sources such as subscales might emerge. No clear standards were used to assess the relative magnitude of the percentage of variance contributed by each source (Schumacker, 2010). Standard errors for variance components provided information about sampling variability of estimated variance components (Webb & Shavelson, 2005).

In a g study analysis, no clear guidelines for the interpretation of results exist (Briesch et al., 2014; Setyonugroho, 2017). However, several sources indicated evaluation of the *g* coefficient should align to standard criteria from reliability coefficients such as Cronbach's alpha. To guide my decision regarding the coefficient, I followed Schumacker (2010) who wrote that for important decisions derived from test scores, a minimum of .90 was appropriate and .95 was preferable. Schumacker argued that .80 or lower was too low when making decisions regarding services and supports for students. Results of the current study were evaluated in relation to the study purpose and the identified sources of variance (Briesch et al., 2014).

Decision Studies

Decision study (d study) designs aim to minimize error for a particular purpose (Shavelson et al., 1989). It is crucial for researchers to conduct a d study after the completion of the g study as the assessment administration conditions leading to the highest levels of reliable and valid scores of the trait being measured is the intention of the statistical model when applied to behavior assessments (Brennan, 2010). For this study, multiple d studies were conducted to investigate rater differences on the same construct of interest with the same student and the interaction effects of rater, item, and person investigating reliability measures for strengths-based behavioral assessments. An important aspect of the d studies that follow this g study are decisions regarding whether the outcomes are applicable to relative decisions or absolute decisions. Both relative *d* and absolute *g* coefficients were included in the d study analysis.

Reliability Scores and Score Interpretation

Reliability scores range from 0 to 1. A score closer to one than to zero is more reliable than scores closer to zero (Brennan, 2010). If the error score is large in relation to the true score, the result is considered inconsistent or unreliable. How scores such as .90 or .70 are interpreted relies upon the statistical lens the researcher uses as well as the parsing of variables into random variables or non-random variables as no clear standards are used to assess the relative magnitude of the percentage of variance contributed by each source (Briesch et al., 2014; Schumacker, 2010; Setyonugroho, 2017). However, if the error score is small in relation to the true score, the result is consistent, reliable, and dependable. Reliability scores less than .70 are considered unsatisfactory, indicating the test items are a not a reliable representation of the construct under measurement (Brennan, 2001). Several sources indicated evaluation of the *g* coefficient should align to standard criteria from reliability coefficients such as Cronbach's alpha (Schumacker, 2010). To guide my decision regarding the coefficient, I followed Schumacker (2010) who wrote that for important decisions derived from test scores, a minimum of .90 was appropriate and .95 was preferable. Furthermore, it was argued that in an applied setting, an even higher threshold should be considered; "a reliability of .90 is the minimum that should be tolerated, and a reliability of .95 should be considered the desirable standard" due to the high stakes decisions being made (Nunnally, 1978, p. 246). While a coefficient alpha of .80 might be considered acceptable when conducting a study in a fully randomized controlled setting, Schumacker argued that .80 or lower was too low when making decisions regarding services and supports for students. Results of the current study were evaluated in relation to the study purpose and the identified sources of variance (Briesch et al., 2014).

CHAPTER IV

RESULTS

Findings

The primary investigation of this study centered on rater variance using the BERS–3 (Epstein & Pierce, in press), a strengths-based assessment. This g study was followed by a d study for the purpose of uncovering the optimal measurement conditions for administering the BERS–3 to clarify and describe conditions where the results might generalize. Secondary data obtained from PRO-ED (2020) were screened for missing values. The BERS–3 was numbered and corresponded to each student represented. If the numbered identification did not have three types of raters associated with it, the case was eliminated from analysis. There were 25 cases corresponding to three rater types. However, there were missing data within some of these cases. This left 22 cases with three complete rater forms corresponding to one matched student for inclusion and analysis in this study. The 22 matched samples met the case or participant threshold set by Webb et al. (1988). Subsequently, data points for the study were reduced to the formula, 22 for *p* x 3 for *r* x 6 categories for *i*, resulting in 396 total data points. A larger set of data points might have provided a more robust analysis. Nevertheless, it was important to note the data points in this study adhered to Webb et al.'s recommendations for g study and subsequent d study analyses.

All data were transferred from Excel to SPSS v.25 for initial scoring. Descriptive statistics were calculated using SPSS v.25 and are provided in Table 8. Assumptions applied to parametric techniques, such as ANOVA, were explored. The level of measurement used a dependent variable at the interval level on a continuous scale. Measurements were not collected in a group setting and there were no interactions among the students or teachers involved in the measurement. Following descriptive statistics, normality was assessed using the explore frequencies function of SPSS v. 25. Using SPSS v. 25 missing data in the 25 matched cases was addressed by excluding cases pairwise. There were four incomplete matched cases, although only three were disqualified from the study due to a violation of the requirements from the test creators regarding missing data. The explore option under the Descriptives tab in SPSS v. 25 was used to gather descriptive data for complete BERS–3 with three corresponding ratings per student rated by teacher, parent, and the student themselves.

Table 8

Descriptive Statistics

Rater Type		Interpersonal Strength	Family Involvement	Intrapersonal Strength	School Functioning	Affective Strengths	Strength Index
Parent	*M*	26.96	20.88	22.92	16.42	14.16	101.79
	SD	8.72	6.17	5.96	5.90	3.10	26.62
	SEM	1.78	1.23	1.19	1.20	.62	5.43
	Var.	76.04	38.03	35.58	34.78	9.64	708.87
	Range	30.00	24.00	25.00	20.00	14.00	100.00
Student	*M*	31.42	21.22	26.00	20.50	13.84	108.35
	SD	7.73	5.00	15.15	5.10	4.30	27.77
	SEM	1.58	1.04	1.05	1.04	.86	5.79
	Var.	59.73	25.097	26.52	26.00	18.56	771.42
	Range	30.00	21.00	19.00	16.00	17.00	122.00
Teacher	*M*	27.08	21.24	21.48	16.64	13.64	100.08
	SD	8.23	4.75	4.63	6.13	3.30	22.81
	SEM	1.64	.95	.93	1.23	.66	4.56
	Var.	67.74	22.53	21.43	37.57	10.91	520.16
	Range	31.00	17.00	19.00	21	15.00	89.00
Total	*M*	86.65	63.26	70.50	54.00	41.64	309.77
	SD	20.10	13.00	11.88	13.18	7.76	55.71
	SEM	4.19	2.71	2.42	2.75	1.55	11.88
	Var.	403.87	168.75	141.13	173.64	60.16	3103.99
	Range	63.00	51.00	47.00	48.00	29.00	197

Note: All *N* are 25 except the following: (a) Parent rater form for Interpersonal Strengths, School Functioning, and Strength Index are 24 and (b) student rater form for Interpersonal Strengths, Intrapersonal Strengths, and School Functioning are 24 and Family Involvement and Strength Index are 23.

Briesch et al. (2014) recommended testing for normality because *SEM* was reported from the g study by calculating the square root of the variance component (absolute or relative); normally distributed data were necessary. Results for all rater types, looking specifically at the BERS-3 (Epstein & Pierce, in press) Strength Index, were *M* = 103.29 with an *SEM* of 3.02, a 95 % confidence interval, an *M* of 97.5796 to 104.2660, *SD* = 25.63764, variance = 657.167, a minimum score of 31 to a maximum score of 152, and a range of 122. Skewness was -.214 with a *SEM* of .283 and Kurtosis was -.077 with an *SEM* of .559. Again, the original mean and the trimmed mean were compared to see if any of the extreme scores had a strong influence on the mean (5% trimmed, *M* = 104.2660). The two scores were far apart, denoting extreme scores significantly impacted the mean. There was a negative skewness (-.214) that showed scores clustered to the right-hand side of the graph. Additionally, a negative kurtosis (-.077) signified the distribution was relatively flat with possible outlier cases; however, it was closer to zero, indicating a better peak than the previous analyses. Figure 2 shows a box plot with outliers. The Kilomogorov-Smirnov test of .050 with a significance value of > .001 indicated no violation of the assumption of normality. A Shapiro-Wilk result of *p* > .001 also meant the data were normally distributed. The distribution of scores was checked using a histogram, which confirmed a normal distribution.

A one-way repeated measure analysis of variance (ANOVA) was conducted to compare scores on the BERS–3 (Epstein & Pierce, in press) Rating Index subscale at each rater level: student, parent, and teacher. The means and standard deviations are presented in Table 9. No significant effect was found for rater: Wilks' Lambda = .929, *F* (2, 20) = .763, *p* > .001, multivariate partial eta squared = .07.

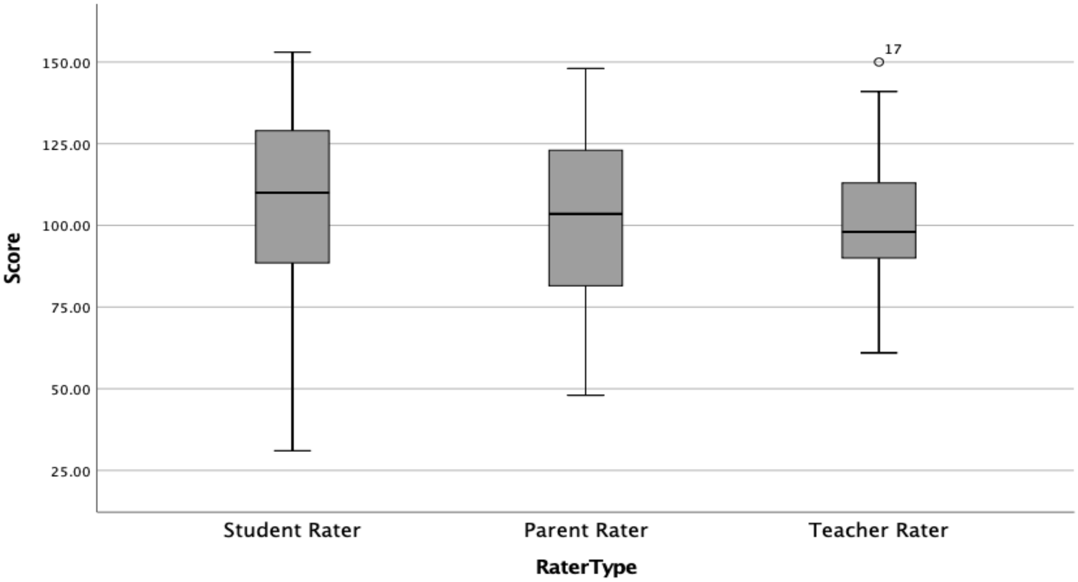


Figure 2. Box plot of three rater types assessing one student.

Tables 9 through 16 provide a closer look at the ranges, means, and standard deviations of the g study inquiry using *person x rater x item* and their interactions. Unlike using SPSS v.25 and CTT, using generalizability theory allowed for all 25 matched cases to be analyzed. The grand mean for the 25 students rated by three rater types using the BERS–3 (Epstein & Pierce, in press) was 34.1 with a variance of 10791, and a standard deviation of 32.9. EduG v.6 allowed for a complete descriptive, statistical analysis for each facet and interaction included in the study.

Table 9

Descriptive Statistics for the Behavioral and Emotional Rating Scale-3 Strength Index Subscale for Student, Parent, and Teacher

Rater Type	*M*	*SD*
Student (1)	108.9091	28.29422
Parent (2)	99.2727	21.04232
Teacher (3)	101.5909	26.76469

N = 22

Table 10

Descriptive Statistics: Participant

Student (p)	*M*	Variance	*SD*
1	24.6	553.0	23.5
2	22.1	389.9	19.7
3	44.8	1719.5	41.5
4	33.2	917.4	30.3
5	31.6	775.8	27.9
6	39.0	1253.2	35.4
7	27.6	638.2	25.3
8	33.8	950.4	30.8
9	33.4	928.8	30.5
10	27.0	485.7	22.0
11	35.8	1135.6	33.7
12	45.9	1734.5	41.6
13	26.2	584.5	24.2
14	41.1	1396.4	37.4
15	39.1	1274.3	35.7
16	41.6	1438.5	37.9
17	37.6	1439.8	37.9
18	34.9	1203.3	34.7
19	38.6	1251.8	35.4
20	23.6	691.6	26.3
21	37.2	1163.7	34.1
22	36.2	1114.8	33.4
23	30.3	750.3	27.4
24	33.2	1176.4	34.3
25	34.1	1029.9	32.1

Table 11

Descriptive Statistics: Rater Type

Rater (r)	*M*	Variance	*SD*
1	33.1	1081.2	32.9
2	33.0	994.5	31.5
3	36.2	1155.3	34.0

Note : Rater type 1=Student; 2=Parent, and 3=Teacher.

Table 12

Descriptive Statistics: Subscale

Subscale (i)	*M*	Variance	*SD*
1	27.9	79.7	8.9
2	21.7	74.1	8.6
3	23.6	30.7	5.5
4	17.8	37.8	6.1
5	13.9	12.4	3.5
6	99.7	964.6	31.1

Note : Subscale 1 = Interpersonal Strength subscale; 2 = Family Involvement subscale; 3= Intrapersonal Strength subscale; 4 = School Functioning subscale; 5 = Affective Strength subscale; 6 = BERS – 3 Composite Index.

Table 13

Descriptive Statistics: Person x Rater Interaction

P Student	R Rater	*M*	Variance	*SD*
1	1	17.3	246.6	15.7
1	2	26.0	554.3	23.5
1	3	30.3	770.6	27.8
2	1	16.0	210.3	14.5
2	2	30.0	732.3	27.1
2	3	20.2	123.8	11.1
3	1	52.8	1950.5	44.2
3	2	30.7	774.2	27.8
3	3	51.0	2131.0	46.2
4	1	29.3	696.6	26.4
4	2	32.7	856.9	29.3
4	3	37.7	1163.6	34.1
5	1	35.3	1025.9	32.0
5	2	22.3	52.9	7.3
5	3	37.0	1119.7	33.5
6	1	39.0	1252.3	35.4
6	2	41.0	1379.7	37.1
6	3	37.0	1119.7	33.5
7	1	24.0	467.7	21.6
7	2	30.0	748.7	27.4
7	3	28.7	678.6	26.0
8	1	34.0	945.7	30.8
8	2	31.0	798.7	28.3
8	3	36.3	1092.6	33.1
9	1	30.3	769.9	27.7
9	2	34.7	967.6	31.1
9	3	35.3	1034.2	32.2
10	1	27.0	597.0	24.4
10	2	21.7	380.6	19.5
10	3	32.3	422.6	20.6
11	1	37.0	1109.0	33.3
11	2	26.3	569.2	23.9
11	3	44.0	1570.3	39.6
12	1	46.0	1721.7	41.5
12	2	47.0	1824.7	42.7
12	3	44.7	1654.6	40.7
13	1	27.0	604.3	24.6
13	2	23.7	463.2	21.5
13	3	28.0	675.7	26.0

Table 13 Continued

P Student	R Rater	*M*	Variance	*SD*
14	1	49.3	1991.6	44.6
14	2	33.7	1015.2	31.9
14	3	40.2	1058.5	32.5
15	1	43.3	1542.6	39.3
15	2	37.3	1145.9	33.9
15	3	36.5	1106.6	33.3
16	1	43.3	1542.6	39.3
16	2	44.7	1633.9	40.4
16	3	36.7	1102.2	33.2
17	1	43.0	1541.7	39.3
17	2	50.0	2037.0	45.1
17	3	19.8	242.1	15.6
18	1	26.3	559.2	23.6
18	2	34.0	1271.7	35.7
18	3	44.3	1615.9	40.2
19	1	33.7	918.6	30.3
19	2	37.7	1162.9	34.1
19	3	44.3	1615.9	40.2
20	1	7.8	66.1	8.1
20	2	31.0	787.7	28.1
20	3	32.0	847.0	29.1
21	1	35.0	1009.3	31.8
21	2	34.7	979.6	31.3
21	3	42.0	1468.0	38.3
22	1	35.0	1009.3	31.8
22	2	31.7	811.6	28.5
22	3	42.0	1468.0	38.3
23	1	27.3	606.9	24.6
23	2	34.0	945.7	30.8
23	3	29.5	675.3	26.0
24	1	30.3	782.2	28.0
24	2	20.3	359.9	19.0
24	3	49.0	1963.7	44.3
25	1	38.3	1199.6	34.6
25	2	38.7	1225.9	35.0
25	3	25.3	548.6	23.4

Note : Rater type 1=Student; 2=Parent, and 3=Teacher.

Table 14

Descriptive Statistics: Person x Item Interaction

P Student	I Subscale	*M*	Variance	*SD*
1	1	20.7	28.2	5.3
1	2	13.3	6.2	2.5
1	3	18.7	24.2	4.9
1	4	11.3	6.2	2.5
1	5	9.7	6.2	2.5
1	6	73.7	262.9	16.2
2	1	20.0	26.0	5.1
2	2	10.7	11.6	3.4
2	3	13.7	17.6	4.2
2	4	15.7	22.9	4.8
2	5	12.0	6.0	2.4
2	6	60.3	444.2	21.1
3	1	33.0	60.7	7.8
3	2	42.7	634.9	25.2
3	3	26.7	48.2	6.9
3	4	20.3	24.2	4.9
3	5	17.7	11.6	3.4
3	6	128.7	696.2	26.4
4	1	24.0	18.7	4.3
4	2	20.0	2.7	1.6
4	3	21.3	17.6	4.2
4	4	19.0	20.7	4.5
4	5	15.3	0.9	0.9
4	6	99.7	105.6	10.3
5	1	31.3	4.2	2.1
5	2	24.3	1.6	1.2
5	3	23.3	5.6	2.4
5	4	20.0	2.7	1.6
5	5	14.0	0.7	0.8
5	6	76.3	2073.6	45.5
6	1	32.7	6.9	2.6
6	2	25.7	1.6	1.2
6	3	23.7	3.6	1.9
6	4	20.7	0.9	0.9
6	5	14.3	0.9	0.9
6	6	117.0	24.0	4.9
7	1	21.7	4.2	2.1
7	2	17.0	8.7	2.9
7	3	20.3	9.6	3.1

Table 14 Continued

P	I	*M*	Variance	*SD*
Student	Subscale			
7	4	9.7	3.6	1.9
7	5	14.0	4.7	2.2
7	6	82.7	59.6	7.7
8	1	29.7	2.9	1.7
8	2	19.3	1.6	1.2
8	3	20.3	27.6	5.2
8	4	19.3	2.9	1.7
8	5	12.7	0.2	0.5
8	6	101.3	42.9	6.5
9	1	24.3	1.6	1.2
9	2	21.3	6.2	2.5
9	3	25.7	6.9	2.6
9	4	13.0	24.7	5.0
9	5	16.0	2.7	1.6
9	6	100.3	44.2	6.6
10	1	20.3	68.2	8.3
10	2	19.0	2.7	1.6
10	3	20.7	33.6	5.8
10	4	13.3	16.2	4.0
10	5	14.3	10.9	3.3
10	6	74.3	46.2	6.8
11	1	26.3	32.9	5.7
11	2	22.7	4.2	2.1
11	3	24.7	24.9	5.0
11	4	17.7	48.2	6.9
11	5	16.0	4.7	2.2
11	6	107.3	474.9	21.8
12	1	40.0	4.7	2.2
12	2	26.7	0.9	0.9
12	3	29.3	1.6	1.2
12	4	24.7	0.9	0.9
12	5	17.0	4.7	2.2
12	6	137.7	8.2	2.9
13	1	23.3	20.2	4.5
13	2	18.0	8.7	2.9
13	3	17.0	6.0	2.4
13	4	12.0	4.7	2.2
13	5	8.3	2.9	1.7
13	6	78.7	30.9	5.6
14	1	40.7	1.6	1.2
14	2	25.0	2.7	1.6
14	3	27.7	46.9	6.8
14	4	18.0	84.7	9.2

Table 14 Continued

P Student	I Subscale	*M*	Variance	*SD*
14	5	15.3	17.6	4.2
14	6	119.7	414.9	20.4
15	1	30.7	8.2	2.9
15	2	25.7	9.6	3.1
15	3	26.7	2.9	1.7
15	4	20.7	10.9	3.3
15	5	13.3	2.9	1.7
15	6	117.3	80.9	9.0
16	1	32.7	17.6	4.2
16	2	27.7	10.9	3.3
16	3	27.3	2.9	1.7
16	4	22.7	1.6	1.2
16	5	14.3	8.2	2.9
16	6	124.7	110.2	10.5
17	1	41.0	0.7	0.8
17	2	19.0	182.0	13.5
17	3	29.0	18.0	4.2
17	4	25.7	0.2	0.5
17	5	18.0	12.7	3.6
17	6	93.0	4398.0	66.3
18	1	21.7	141.6	11.9
18	2	21.3	14.2	3.8
18	3	24.0	26.0	5.1
18	4	19.7	11.6	3.4
18	5	14.3	4.2	2.1
18	6	108.3	496.9	22.3
19	1	31.3	32.9	5.7
19	2	23.0	2.0	1.4
19	3	25.0	18.0	4.2
19	4	21.7	0.9	0.9
19	5	14.7	2.9	1.7
19	6	115.7	174.2	13.2
20	1	16.3	134.9	11.6
20	2	17.7	6.9	2.6
20	3	21.7	2.9	1.7
20	4	11.7	76.2	8.7
20	5	11.3	5.6	2.4
20	6	63.0	1986.0	44.6
21	1	30.7	34.9	5.9
21	2	21.7	3.6	1.9
21	3	27.3	2.9	1.7
21	4	17.7	6.9	2.6
21	5	14.3	0.2	0.5

Table 14 Continued

P	I	*M*	Variance	*SD*
Student	Subscale			
21	6	111.7	102.9	10.1
22	1	29.3	48.2	6.9
22	2	22.3	0.9	0.9
22	3	25.3	20.2	4.5
22	4	17.7	6.9	2.6
22	5	14.0	0.7	0.8
22	6	108.7	166.9	12.9
23	1	20.0	20.7	4.5
23	2	19.0	8.7	2.9
23	3	22.0	10.7	3.3
23	4	22.3	14.9	3.9
23	5	8.7	10.9	3.3
23	6	89.7	77.6	8.8
24	1	25.3	150.9	12.3
24	2	22.3	29.6	5.4
24	3	26.3	22.9	4.8
24	4	12.0	84.7	9.2
24	5	13.7	24.2	4.9
24	6	99.7	1270.2	35.6
25	1	30.3	14.9	3.9
25	2	17.7	48.2	6.9
25	3	21.7	10.9	3.3
25	4	18.7	2.9	1.7
25	5	14.0	18.7	4.3
25	6	102.3	346.9	18.6

Note: Subscale 1 = Interpersonal Strength subscale; 2 = Family Involvement subscale; 3= Intrapersonal Strength subscale; 4 = School Functioning subscale; 5 = Affective Strength subscale; 6 = BERS – 3 Composite Index.

Table 15

Descriptive Statistics: Rater x Item Interaction

R Rater	I Subscale	*M*	Variance	*SD*
1	1	25.3	86.0	9.3
1	2	22.9	161.0	12.7
1	3	22.9	34.2	5.8
1	4	15.8	42.3	6.5
1	5	14.2	9.3	3.0
1	6	97.7	1050.0	32.4
2	1	26.7	74.0	8.6
2	2	21.7	19.0	4.4
2	3	21.4	20.5	4.5
2	4	17.1	34.9	5.9
2	5	13.8	9.9	3.1
2	6	97.2	763.7	27.6
3	1	31.7	56.6	7.5
3	2	20.6	39.8	6.3
3	3	26.4	24.6	5.0
3	4	20.6	23.9	4.9
3	5	13.7	17.8	4.2
3	6	104.1	1050.7	32.4

Note : Rater type 1=Student; 2=Parent, and 3=Teacher. Subscale 1 = Interpersonal Strength subscale; 2 = Family Involvement subscale; 3= Intrapersonal Strength subscale; 4 = School Functioning subscale; 5 = Affective Strength subscale; 6 = BERS – 3 Composite Index.

Table 16

Descriptive Statistics: Person x Rater x Item Interaction

P Student	R Rater	I Subscale	*M*	Variance	*SD*
1	1	1	14.0	0.0	0.0
1	1	2	10.0	0.0	0.0
1	1	3	13.0	0.0	0.0
1	1	4	8.0	0.0	0.0
1	1	5	7.0	0.0	0.0
1	1	6	52.0	0.0	0.0
1	2	1	21.0	0.0	0.0
1	2	2	16.0	0.0	0.0
1	2	3	18.0	0.0	0.0
1	2	4	14.0	0.0	0.0
1	2	5	9.0	0.0	0.0
1	2	6	78.0	0.0	0.0
1	3	1	27.0	0.0	0.0
1	3	2	14.0	0.0	0.0
1	3	3	25.0	0.0	0.0
1	3	4	12.0	0.0	0.0
1	3	5	13.0	0.0	0.0
1	3	6	91.0	0.0	0.0
2	1	1	13.0	0.0	0.0
2	1	2	6.0	0.0	0.0
2	1	3	8.0	0.0	0.0
2	1	4	9.0	0.0	0.0
2	1	5	12.0	0.0	0.0
2	1	6	48.0	0.0	0.0
2	2	1	25.0	0.0	0.0
2	2	2	14.0	0.0	0.0
2	2	3	18.0	0.0	0.0
2	2	4	18.0	0.0	0.0
2	2	5	15.0	0.0	0.0
2	2	6	90.0	0.0	0.0
2	3	1	22.0	0.0	0.0
2	3	2	12.0	0.0	0.0
2	3	3	15.0	0.0	0.0
2	3	4	20.0	0.0	0.0
2	3	5	9.0	0.0	0.0
2	3	6	43.0	0.0	0.0
3	1	1	28.0	0.0	0.0
3	1	2	78.0	0.0	0.0
3	1	3	30.0	0.0	0.0

Table 16 Continued

P Student	R Rater	I Subscale	*M*	Variance	*SD*
3	1	4	21.0	0.0	0.0
3	1	5	19.0	0.0	0.0
3	1	6	141.0	0.0	0.0
3	2	1	27.0	0.0	0.0
3	2	2	21.0	0.0	0.0
3	2	3	17.0	0.0	0.0
3	2	4	14.0	0.0	0.0
3	2	5	13.0	0.0	0.0
3	2	6	92.0	0.0	0.0
3	3	1	44.0	0.0	0.0
3	3	2	29.0	0.0	0.0
3	3	3	33.0	0.0	0.0
3	3	4	26.0	0.0	0.0
3	3	5	21.0	0.0	0.0
3	3	6	153.0	0.0	0.0
4	1	1	20.0	0.0	0.0
4	1	2	22.0	0.0	0.0
4	1	3	17.0	0.0	0.0
4	1	4	13.0	0.0	0.0
4	1	5	16.0	0.0	0.0
4	1	6	88.0	0.0	0.0
4	2	1	22.0	0.0	0.0
4	2	2	20.0	0.0	0.0
4	2	3	20.0	0.0	0.0
4	2	4	20.0	0.0	0.0
4	2	5	16.0	0.0	0.0
4	2	6	98.0	0.0	0.0
4	3	1	30.0	0.0	0.0
4	3	2	18.0	0.0	0.0
4	3	3	27.0	0.0	0.0
4	3	4	24.0	0.0	0.0
4	3	5	14.0	0.0	0.0
4	3	6	113.0	0.0	0.0
5	1	1	31.0	0.0	0.0
5	1	2	23.0	0.0	0.0
5	1	3	20.0	0.0	0.0
5	1	4	18.0	0.0	0.0
5	1	5	14.0	0.0	0.0
5	1	6	106.0	0.0	0.0
5	2	1	34.0	0.0	0.0
5	2	2	26.0	0.0	0.0
5	2	3	25.0	0.0	0.0
5	2	4	22.0	0.0	0.0

Table 16 Continued

P Student	R Rater	I Subscale	*M*	Variance	*SD*
5	2	5	15.0	0.0	0.0
5	2	6	12.0	0.0	0.0
5	3	1	29.0	0.0	0.0
5	3	2	24.0	0.0	0.0
5	3	3	25.0	0.0	0.0
5	3	4	20.0	0.0	0.0
5	3	5	13.0	0.0	0.0
5	3	6	111.0	0.0	0.0
6	1	1	34.0	0.0	0.0
6	1	2	27.0	0.0	0.0
6	1	3	21.0	0.0	0.0
6	1	4	20.0	0.0	0.0
6	1	5	15.0	0.0	0.0
6	1	6	117.0	0.0	0.0
6	2	1	35.0	0.0	0.0
6	2	2	26.0	0.0	0.0
6	2	3	25.0	0.0	0.0
6	2	4	22.0	0.0	0.0
6	2	5	15.0	0.0	0.0
6	2	6	123.0	0.0	0.0
6	3	1	29.0	0.0	0.0
6	3	2	24.0	0.0	0.0
6	3	3	25.0	0.0	0.0
6	3	4	20.0	0.0	0.0
6	3	5	13.0	0.0	0.0
6	3	6	111.0	0.0	0.0
7	1	1	19.0	0.0	0.0
7	1	2	14.0	0.0	0.0
7	1	3	16.0	0.0	0.0
7	1	4	11.0	0.0	0.0
7	1	5	12.0	0.0	0.0
7	1	6	72.0	0.0	0.0
7	2	1	22.0	0.0	0.0
7	2	2	21.0	0.0	0.0
7	2	3	23.0	0.0	0.0
7	2	4	7.0	0.0	0.0
7	2	5	17.0	0.0	0.0
7	2	6	90.0	0.0	0.0
7	3	1	24.0	0.0	0.0
7	3	2	16.0	0.0	0.0
7	3	3	22.0	0.0	0.0
7	3	4	11.0	0.0	0.0
7	3	5	13.0	0.0	0.0

Table 16 Continued

P Student	R Rater	I Subscale	*M*	Variance	*SD*
7	3	6	86.0	0.0	0.0
8	1	1	28.0	0.0	0.0
8	1	2	18.0	0.0	0.0
8	1	3	23.0	0.0	0.0
8	1	4	20.0	0.0	0.0
8	1	5	13.0	0.0	0.0
8	1	6	102.0	0.0	0.0
8	2	1	29.0	0.0	0.0
8	2	2	21.0	0.0	0.0
8	2	3	13.0	0.0	0.0
8	2	4	17.0	0.0	0.0
8	2	5	13.0	0.0	0.0
8	2	6	93.0	0.0	0.0
8	3	1	32.0	0.0	0.0
8	3	2	19.0	0.0	0.0
8	3	3	25.0	0.0	0.0
8	3	4	21.0	0.0	0.0
8	3	5	12.0	0.0	0.0
8	3	6	109.0	0.0	0.0
9	1	1	23.0	0.0	0.0
9	1	2	18.0	0.0	0.0
9	1	3	27.0	0.0	0.0
9	1	4	9.0	0.0	0.0
9	1	5	14.0	0.0	0.0
9	1	6	91.0	0.0	0.0
9	2	1	24.0	0.0	0.0
9	2	2	22.0	0.0	0.0
9	2	3	22.0	0.0	0.0
9	2	4	20.0	0.0	0.0
9	2	5	16.0	0.0	0.0
9	2	6	104.0	0.0	0.0
9	3	1	26.0	0.0	0.0
9	3	2	24.0	0.0	0.0
9	3	3	28.0	0.0	0.0
9	3	4	10.0	0.0	0.0
9	3	5	18.0	0.0	0.0
9	3	6	106.0	0.0	0.0
10	1	1	15.0	0.0	0.0
10	1	2	19.0	0.0	0.0
10	1	3	22.0	0.0	0.0
10	1	4	10.0	0.0	0.0
10	1	5	15.0	0.0	0.0
10	1	6	81.0	0.0	0.0

Table 16 Continued

P Student	R Rater	I Subscale	*M*	Variance	*SD*
10	2	1	14.0	0.0	0.0
10	2	2	17.0	0.0	0.0
10	2	3	13.0	0.0	0.0
10	2	4	11.0	0.0	0.0
10	2	5	10.0	0.0	0.0
10	2	6	65.0	0.0	0.0
10	3	1	32.0	0.0	0.0
10	3	2	21.0	0.0	0.0
10	3	3	27.0	0.0	0.0
10	3	4	19.0	0.0	0.0
10	3	5	18.0	0.0	0.0
10	3	6	77.0	0.0	0.0
11	1	1	27.0	0.0	0.0
11	1	2	23.0	0.0	0.0
11	1	3	26.0	0.0	0.0
11	1	4	18.0	0.0	0.0
11	1	5	17.0	0.0	0.0
11	1	6	111.0	0.0	0.0
11	2	1	19.0	0.0	0.0
11	2	2	20.0	0.0	0.0
11	2	3	18.0	0.0	0.0
11	2	4	9.0	0.0	0.0
11	2	5	13.0	0.0	0.0
11	2	6	79.0	0.0	0.0
11	3	1	33.0	0.0	0.0
11	3	2	25.0	0.0	0.0
11	3	3	30.0	0.0	0.0
11	3	4	26.0	0.0	0.0
11	3	5	18.0	0.0	0.0
11	3	6	132.0	0.0	0.0
12	1	1	37.0	0.0	0.0
12	1	2	26.0	0.0	0.0
12	1	3	31.0	0.0	0.0
12	1	4	24.0	0.0	0.0
12	1	5	20.0	0.0	0.0
12	1	6	138.0	0.0	0.0
12	2	1	42.0	0.0	0.0
12	2	2	28.0	0.0	0.0
12	2	3	29.0	0.0	0.0
12	2	4	26.0	0.0	0.0
12	2	5	16.0	0.0	0.0
12	2	6	141.0	0.0	0.0
12	3	1	41.0	0.0	0.0

Table 16 Continued

P Student	R Rater	I Subscale	*M*	Variance	*SD*
12	3	2	26.0	0.0	0.0
12	3	3	28.0	0.0	0.0
12	3	4	24.0	0.0	0.0
12	3	5	15.0	0.0	0.0
12	3	6	134.0	0.0	0.0
13	1	1	23.0	0.0	0.0
13	1	2	17.0	0.0	0.0
13	1	3	20.0	0.0	0.0
13	1	4	11.0	0.0	0.0
13	1	5	10.0	0.0	0.0
13	1	6	81.0	0.0	0.0
13	2	1	18.0	0.0	0.0
13	2	2	15.0	0.0	0.0
13	2	3	17.0	0.0	0.0
13	2	4	15.0	0.0	0.0
13	2	5	6.0	0.0	0.0
13	2	6	71.0	0.0	0.0
13	3	1	29.0	0.0	0.0
13	3	2	22.0	0.0	0.0
13	3	3	14.0	0.0	0.0
13	3	4	10.0	0.0	0.0
13	3	5	9.0	0.0	0.0
13	3	6	84.0	0.0	0.0
14	1	1	42.0	0.0	0.0
14	1	2	27.0	0.0	0.0
14	1	3	33.0	0.0	0.0
14	1	4	25.0	0.0	0.0
14	1	5	21.0	0.0	0.0
14	1	6	148.0	0.0	0.0
14	2	1	39.0	0.0	0.0
14	2	2	25.0	0.0	0.0
14	2	3	18.0	0.0	0.0
14	2	4	5.0	0.0	0.0
14	2	5	14.0	0.0	0.0
14	2	6	101.0	0.0	0.0
14	3	1	41.0	0.0	0.0
14	3	2	23.0	0.0	0.0
14	3	3	32.0	0.0	0.0
14	3	4	24.0	0.0	0.0
14	3	5	11.0	0.0	0.0
14	3	6	110.0	0.0	0.0
15	1	1	34.0	0.0	0.0
15	1	2	30.0	0.0	0.0

Table 16 Continued

P Student	R Rater	I Subscale	*M*	Variance	*SD*
15	1	3	29.0	0.0	0.0
15	1	4	23.0	0.0	0.0
15	1	5	14.0	0.0	0.0
15	1	6	130.0	0.0	0.0
15	2	1	31.0	0.0	0.0
15	2	2	24.0	0.0	0.0
15	2	3	26.0	0.0	0.0
15	2	4	16.0	0.0	0.0
15	2	5	15.0	0.0	0.0
15	2	6	112.0	0.0	0.0
15	3	1	27.0	0.0	0.0
15	3	2	23.0	0.0	0.0
15	3	3	25.0	0.0	0.0
15	3	4	23.0	0.0	0.0
15	3	5	11.0	0.0	0.0
15	3	6	110.0	0.0	0.0
16	1	1	34.0	0.0	0.0
16	1	2	30.0	0.0	0.0
16	1	3	29.0	0.0	0.0
16	1	4	23.0	0.0	0.0
16	1	5	14.0	0.0	0.0
16	1	6	130.0	0.0	0.0
16	2	1	37.0	0.0	0.0
16	2	2	30.0	0.0	0.0
16	2	3	28.0	0.0	0.0
16	2	4	21.0	0.0	0.0
16	2	5	18.0	0.0	0.0
16	2	6	134.0	0.0	0.0
16	3	1	27.0	0.0	0.0
16	3	2	23.0	0.0	0.0
16	3	3	25.0	0.0	0.0
16	3	4	24.0	0.0	0.0
16	3	5	11.0	0.0	0.0
16	3	6	110.0	0.0	0.0
17	1	1	40.0	0.0	0.0
17	1	2	27.0	0.0	0.0
17	1	3	23.0	0.0	0.0
17	1	4	26.0	0.0	0.0
17	1	5	13.0	0.0	0.0
17	1	6	129.0	0.0	0.0
17	2	1	41.0	0.0	0.0
17	2	2	30.0	0.0	0.0
17	2	3	32.0	0.0	0.0

Table 16 Continued

P Student	R Rater	I Subscale	*M*	Variance	*SD*
17	2	4	26.0	0.0	0.0
17	2	5	21.0	0.0	0.0
17	2	6	150.0	0.0	0.0
17	3	1	42.0	0.0	0.0
17	3	2	0.0	0.0	0.0
17	3	3	32.0	0.0	0.0
17	3	4	25.0	0.0	0.0
17	3	5	20.0	0.0	0.0
17	3	6	0.0	0.0	0.0
18	1	1	17.0	0.0	0.0
18	1	2	16.0	0.0	0.0
18	1	3	19.0	0.0	0.0
18	1	4	15.0	0.0	0.0
18	1	5	12.0	0.0	0.0
18	1	6	79.0	0.0	0.0
18	2	1	10.0	0.0	0.0
18	2	2	24.0	0.0	0.0
18	2	3	22.0	0.0	0.0
18	2	4	21.0	0.0	0.0
18	2	5	14.0	0.0	0.0
18	2	6	113.0	0.0	0.0
18	3	1	38.0	0.0	0.0
18	3	2	24.0	0.0	0.0
18	3	3	31.0	0.0	0.0
18	3	4	23.0	0.0	0.0
18	3	5	17.0	0.0	0.0
18	3	6	133.0	0.0	0.0
19	1	1	24.0	0.0	0.0
19	1	2	21.0	0.0	0.0
19	1	3	22.0	0.0	0.0
19	1	4	21.0	0.0	0.0
19	1	5	13.0	0.0	0.0
19	1	6	101.0	0.0	0.0
19	2	1	32.0	0.0	0.0
19	2	2	24.0	0.0	0.0
19	2	3	22.0	0.0	0.0
19	2	4	21.0	0.0	0.0
19	2	5	14.0	0.0	0.0
19	2	6	113.0	0.0	0.0
19	3	1	38.0	0.0	0.0
19	3	2	24.0	0.0	0.0
19	3	3	31.0	0.0	0.0
19	3	4	23.0	0.0	0.0

Table 16 Continued

P Student	R Rater	I Subscale	*M*	Variance	*SD*
19	3	5	17.0	0.0	0.0
19	3	6	133.0	0.0	0.0
20	1	1	0.0	0.0	0.0
20	1	2	14.0	0.0	0.0
20	1	3	20.0	0.0	0.0
20	1	4	0.0	0.0	0.0
20	1	5	13.0	0.0	0.0
20	1	6	0.0	0.0	0.0
20	2	1	26.0	0.0	0.0
20	2	2	19.0	0.0	0.0
20	2	3	21.0	0.0	0.0
20	2	4	14.0	0.0	0.0
20	2	5	13.0	0.0	0.0
20	2	6	93.0	0.0	0.0
20	3	1	23.0	0.0	0.0
20	3	2	20.0	0.0	0.0
20	3	3	24.0	0.0	0.0
20	3	4	21.0	0.0	0.0
20	3	5	8.0	0.0	0.0
20	3	6	96.0	0.0	0.0
21	1	1	26.0	0.0	0.0
21	1	2	23.0	0.0	0.0
21	1	3	28.0	0.0	0.0
21	1	4	14.0	0.0	0.0
21	1	5	14.0	0.0	0.0
21	1	6	105.0	0.0	0.0
21	2	1	27.0	0.0	0.0
21	2	2	19.0	0.0	0.0
21	2	3	25.0	0.0	0.0
21	2	4	19.0	0.0	0.0
21	2	5	14.0	0.0	0.0
21	2	6	104.0	0.0	0.0
21	3	1	39.0	0.0	0.0
21	3	2	23.0	0.0	0.0
21	3	3	29.0	0.0	0.0
21	3	4	20.0	0.0	0.0
21	3	5	15.0	0.0	0.0
21	3	6	126.0	0.0	0.0
22	1	1	26.0	0.0	0.0
22	1	2	23.0	0.0	0.0
22	1	3	28.0	0.0	0.0
22	1	4	14.0	0.0	0.0
22	1	5	14.0	0.0	0.0

Table 16 Continued

P Student	R Rater	I Subscale	*M*	Variance	*SD*
22	1	6	105.0	0.0	0.0
22	2	1	23.0	0.0	0.0
22	2	2	21.0	0.0	0.0
22	2	3	19.0	0.0	0.0
22	2	4	19.0	0.0	0.0
22	2	5	13.0	0.0	0.0
22	2	6	95.0	0.0	0.0
22	3	1	39.0	0.0	0.0
22	3	2	23.0	0.0	0.0
22	3	3	29.0	0.0	0.0
22	3	4	20.0	0.0	0.0
22	3	5	15.0	0.0	0.0
22	3	6	126.0	0.0	0.0
23	1	1	21.0	0.0	0.0
23	1	2	15.0	0.0	0.0
23	1	3	18.0	0.0	0.0
23	1	4	17.0	0.0	0.0
23	1	5	11.0	0.0	0.0
23	1	6	82.0	0.0	0.0
23	2	1	25.0	0.0	0.0
23	2	2	20.0	0.0	0.0
23	2	3	22.0	0.0	0.0
23	2	4	24.0	0.0	0.0
23	2	5	11.0	0.0	0.0
23	2	6	102.0	0.0	0.0
23	3	1	14.0	0.0	0.0
23	3	2	22.0	0.0	0.0
23	3	3	26.0	0.0	0.0
23	3	4	26.0	0.0	0.0
23	3	5	4.0	0.0	0.0
23	3	6	85.0	0.0	0.0
24	1	1	24.0	0.0	0.0
24	1	2	24.0	0.0	0.0
24	1	3	24.0	0.0	0.0
24	1	4	6.0	0.0	0.0
24	1	5	13.0	0.0	0.0
24	1	6	91.0	0.0	0.0
24	2	1	11.0	0.0	0.0
24	2	2	15.0	0.0	0.0
24	2	3	22.0	0.0	0.0
24	2	4	5.0	0.0	0.0
24	2	5	8.0	0.0	0.0
24	2	6	61.0	0.0	0.0

Table 16 Continued

P Student	R Rater	I Subscale	*M*	Variance	*SD*
24	3	1	41.0	0.0	0.0
24	3	2	28.0	0.0	0.0
24	3	3	33.0	0.0	0.0
24	3	4	25.0	0.0	0.0
24	3	5	20.0	0.0	0.0
24	3	6	147.0	0.0	0.0
25	1	1	32.0	0.0	0.0
25	1	2	21.0	0.0	0.0
25	1	3	26.0	0.0	0.0
25	1	4	18.0	0.0	0.0
25	1	5	18.0	0.0	0.0
25	1	6	115.0	0.0	0.0
25	2	1	34.0	0.0	0.0
25	2	2	24.0	0.0	0.0
25	2	3	21.0	0.0	0.0
25	2	4	21.0	0.0	0.0
25	2	5	16.0	0.0	0.0
25	2	6	116.0	0.0	0.0
25	3	1	25.0	0.0	0.0
25	3	2	8.0	0.0	0.0
25	3	3	18.0	0.0	0.0
25	3	4	17.0	0.0	0.0
25	3	5	8.0	0.0	0.0
25	3	6	76.0	0.0	0.0

Note: Rater type 1=Student; 2=Parent, and 3=Teacher. Subscale 1 = Interpersonal Strength subscale; 2 = Family Involvement subscale; 3= Intrapersonal Strength subscale; 4 = School Functioning subscale; 5 = Affective Strength subscale; 6 = BERS – 3 Composite Index.

To test for homogeneity of variance, Levene's test for equality was computed. For this test, I looked at each subscale separately to evaluate any variation between rater types: (a) Interpersonal Strength (*p* = .85), (b) Family Involvement (*p* = .00), (c) Intrapersonal Strength (*p* = .21), (d) School Functioning *(p* = .20), (e) Affective Strength (*p* = .04), and (f) Strength Index (*p* = .68). Scores from the Family Involvement and Affective Strength subscales indicated the variance between the three groups was not equal and the assumption of homogeneity of variance was violated. For ANOVA and because the size of the groups was almost identical, it did not cause a problem to continue to run that test (Pallant, 2016). However, it was important to note the Welch and Brown-Forsythe Robust Tests of Equality of Means had a *p* value of .573 and .646, respectively, when analyzing the Affective Strength subscale, indicating the assumption of homogeneity of variance was not violated. The Family Involvement subscale had *p* values of .018 and .003 for the Welch and Brown-Forsythe Robust Tests of Equality of Means, respectively; the assumption of homogeneity of variance continued to be violated for this particular subscale of the BERS-3 (Epstein & Pierce, in press).

An ANOVA was conducted using the general linear model function in SPSS v. 25 to compare scores on the subscales for the parent, student, and teacher raters. The results are found in Table 17. EduG v.6 also produced results for ANOVA as part of the G theory analysis. A statistically significant difference was found at the $p < .05$ level in the Family Involvement, $F(2, 69) = 6.38$, $p = .003$; Intrapersonal Strength, $F(2, 70) = 20.10$, $p = .000$; and School Functioning subscales, $F(2, 70) = 4.41$, $p = .016$ for the three rater types.

Table 17

Analysis of Variance Using Statistical Package for the Social Sciences Version 25

		Sum of Squares	df	Mean Square	*F*	*p*	Eta squared
Interpersonal Strength	Between Groups	311.53	2	155.77	2.30	.108	.06
	Within Groups	4748.63	70	67.84			
	Total	5060.16	72				
Family Involvement	Between Groups	525.18	2	262.60	6.38	.003	.16
	Within Groups	2841.43	69	41.18			
	Total	3366.61	71				
Intrapersonal Strength	Between Groups	1103.30	2	551.65	20.10	<.001	.36
	Within Groups	1924.07	70	27.49			
	Total	3027.27	72				
School Functioning	Between Groups	281.78	2	140.89	4.41	.016	.11
	Within Groups	2238.55	70	31.98			
	Total	2520.33	72				
Affective Strength	Between Groups	11.417	2	5.708	.436	.648	.01
	Within Groups	929.72	71	13.095			
	Total	941.14	73				
BERS – 3 Strength Index	Between Groups	899.86	2	449.93	.678	.511	.02
	Within Groups	45759.02	69	663.17			
	Total	46658.79	71				

Upon further analysis, it was found that using the multiple comparisons Tukey HSD post-hoc feature in SPSS v. 25 on the Family Involvement subscale, the student and parent ratings (*M* = 21.22, *SD* =5.01 and *M* =26.96, *SD* = 8.72, respectively) and teacher and parent ratings (*M* = 21.24, *SD* =4.75 and *M* =26.96, *SD* = 8.72, respectively) were statistically significant. The Intrapersonal Strength subscale showed statistical significance between the student and parent (*M* = 26.00, *SD* =5.15 and *M* =16.42, *SD* = 5.90, respectively), student and teacher (*M* = 26.00, *SD* =5.15 and *M* =21.48, *SD* = 4.63, respectively), and parent and teacher (*M* =16.42, *SD* = 5.90 and *M* =21.48, *SD* = 4.63, respectively) ratings. Finally, on the School Functioning subscale, the student and parent (*M* =20.71, *SD* = 4.83 and *M* =14.16, *SD* = 3.10, respectively) and student and teacher (*M* =20.71, *SD* = 4.83 and *M* =13.64, *SD* = 3.30, respectively) ratings were statistically significant. The means plots for this ANOVA indicated students rated themselves much higher than the other two raters on the Interpersonal Strength, Intrapersonal Strength, School Functioning, and BERS-3 (Epstein & Pierce, in press) Strength Index. Parent ratings were highest for the Affective Strength subscale. Next, eta squared was calculated for effect size. Small effect size was found for the Affective Strength subscale (.01) and the BERS-3 Strength Index (.02). A medium effect size for the Interpersonal Strength subscale (.06), large effect sizes for the Family Involvement (.16) and School Functioning subscales (.11), and a very large effect size for the Intrapersonal Strength subscale (.36) were found.

Generalizability Study

The g study design used to answer the two research questions was *p* (student) x *r* (rater type: student, parent, and teacher) x *i* (subscale). The proportion of rater variance in the 25 cases using the BERS-3 (Epstein & Pierce, in press) was calculated for all subscales as well as the BERS-3 Strength Index. This g study design, *p* x *r* x *i*, followed recommendations for g study parameters set forth by Brennan (1992) and Shavelson and Webb (1991). Tables 18 through 22 represent the results derived from EduG v. 6.

Table 18

Generalizability Study Observation and Estimation Design for p x r x i

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
Student	P	23	INF	
Rater	R	3	INF	
Subscale	I	6	INF	

Table 19

Analysis of Variance Using EduG v. 6 for the p x r x i Study Design

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	15668.6	22	712.2	21.5	21.5	21.5	1.7	11.8
R	1580.3	2	790.1	3.3	3.3	3.3	0.3	4.1
I	381134.6	5	76226.9	1100.9	1100.9	1100.9	87.3	590.5
PR	10476.6	44	238.1	27.2	27.2	27.2	2.2	8.4
PI	17851.6	110	162.3	29.1	29.1	29.1	2.3	7.6
RI	1764.3	10	176.4	4.4	4.4	4.4	0.3	3.1
PRI	16486.8	220	74.9	74.9	74.9	74.9	5.9	7.1
Total	444962.9	413					100	

Table 20

Generalizability Study Analysis using EduG v. 6 for the p x r x I Study Design

Source of Variance	Differentiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	21.5		.....		.....	
	.....	R	.....		1.1	0.5
	.....	I	.....		183.5	90.4
	.....	PR	9.1	50.1	9.1	4.5
	.....	PI	4.9	26.8	4.9	2.4
	.....	RI	.....		0.2	0.1
	.....	PRI	4.2	23.0	4.2	2.1
Sum of variances	21.5		18.1	100.0	202.9	100.0
SD	4.6		Relative SE : 4.3		Absolute SE :14.2	
Relative Coefficient *G*	0.54					
Absolute Coefficient *G*	0.10					

Grand mean for levels used: 34.4
Variance error of the mean for levels used: 186.5
Standard error of the grand mean: 13.7

Variance error of the mean for levels used was 186.5. This represented the mean of all acceptable observations. The standard error of the grand mean was 34.4. This was the amount the sample population mean would differ from the true population mean. The relative *g* coefficient was .54 and the absolute *g* coefficient was .10. The estimated proportion of the observed score variance due to universe score variation was .54

(analogous to the true score variance in CTT). To evaluate these results, it was important to remember that each set of variance components was unique in relation to the study purpose and the identified sources of variance. Additionally, the difference between relative sources of variance and absolute sources of variance must be deconstructed. The relative source of variance represented by the *g* coefficient (.54) was akin to the reliability coefficient in CTT and might be interpreted using Cronbach's alpha (Lakes & Hoyt, 2009; Naumenko, 2015). The relative source of variance allowed for a norm-referenced interpretation of test results (e.g., a student's scores in comparison to other students' scores). Sometimes called the *d* coefficient or the dependability coefficient, the absolute *g* coefficient represented a criterion-referenced interpretation of test results. This is where there was a fixed cut score students were measured against. Results from this g study indicated the interaction of rater type and student variance accounted for almost 50% of the error or variance for the BERS-3 (Epstein & Pierce, in press). This large variance could happen when the rater rated a particular student more favorably than another rater did despite the fact they tended to agree on other ratings. Using the *d* coefficient criterion, a result of .10 told us that if we were making decisions regarding supports and services for students, the type of rater mattered.

CHAPTER V

DISCUSSION AND CONCLUSION

By uncovering more information about cross-informant characteristics that influence the outcomes derived from data retrieved from a strengths-based behavior rating scale for emotional and behavioral disorders, educators could increase the accuracy and appropriateness of identification for and provision of services for students with emotional disturbance. Using a strengths-based rating scale might function not only as a proactive approach to addressing student need but would also provide valuable information to address changes in student needs moving forward (Bruhn et al., 2014). Behavioral rating scales are among the most commonly used measures to identify at-risk students (Briesch et al., 2016). Furthermore, they are also used as markers for monitoring social, emotional, and behavioral fluctuations in students over time, aid in the distribution of educational resources and services best suited for the individual needs of the student, and increase the communication and collaboration among members of the IEP team and the student’s family (Duppong Hurley et al., 2015; Epstein et al., 2000; Kaurin et al., 2016). Understanding factors affecting students’ scores on behavior rating scales resultant from different types of raters might inform rater training and selection of raters, leading to more accurate information for decisions about special education services, mental health support, or other accommodations necessary to provide FAPE.

The primary purpose of this study was to evaluate, using generalizability theory, potential variances between types of raters using a strengths-based assessment to measure the emotional and behavioral domains of students in middle and high school. The greater aim was to gather this level of information to advise administration procedures and informant training so decisions derived from strengths-based behavior rating scales were accurate and appropriate. If our methods for data collection are rife with inadequacy and error, a barrier is created for students' access to FAPE. Often behavior rating scales are among the first type of assessment component in a multi-tiered system of support giving educators information to focus on a smaller population of interest with special characteristics under evaluation (Bruhn et al., 2014). These measures are used to identify factors that contribute to negative outcomes and an increased inability to learn. The importance of ensuring collected data from this assessment is irrefutable and cannot be expressed enough.

This g study was designed to answer the following research questions:

Q1 To what extent does rater type (e.g., student, parent, and teacher) explain rater scores derived from a strengths-based behavioral rating scale?

Q2 To what extent are the scores derived from three rater informants on a strengths-based behavioral rating scale reliable for use in absolute or relative decisions?

The results from the data analysis procedures are discussed in this chapter. Additionally, two new g study designs controlling for the variable of gender, *g* facet analysis, and subsequent *d* study optimization are presented with implications discussed.

Research Question Results

Results from the both the descriptive statistics and ANOVA suggested ratings obtained from the BERS-3 (Epstein & Pierce, in press) Parent Rating Scale, Youth Rating Scale, and Teacher Rating Scale were reliable measures of the construct of interest among multiple informants. When examining *p* values to determine statistical significance, the Family Involvement, Intrapersonal Strengths, and School Functioning subscales all had *p* values indicating variation was unlikely due to chance. However, I followed this by examining effect sizes using partial eta-squared that flagged a medium effect size for the Interpersonal Strength subscale (.06), a large effect size for the Family Involvement (.16) and School Functioning subscales (.11), and a very large effect size for the Intrapersonal Strength subscale (.36). Family Involvement, School Functioning, and Intrapersonal Strengths subscale scores were found to be both statistically significant and had medium to very large effect sizes. The student rater informant reported significantly better results on these three subscales than the teacher or parent informants. This could indicate the students' perceptions of their capabilities, their place within the family structure, and their ability to do well in school were better than they actually were. Positive illusory bias, an "overly positive view of oneself despite contradictory external indices to [the] contrary," is a well-researched phenomenon impacting the accuracy of self-report among students with ED (Gage & Lierheimer, 2012, p. 2; Volz-Sidiropoulou, Boecker, & Gauggel, 2016). Results might indicate that when evaluating informant reports for behavioral rating scales, the results produced from the student themselves should be viewed with caution. Self-reporting might be dubious. Many studies agreed that students with ED tended to rate themselves more favorably than where their skills, attributes, and competencies levels truly were (Volz-Sidiropoulou et al., 2016). However, the information provided by student self-report in and of itself might help parents and teachers learn more about where students felt confident about themselves and how to draw upon those areas where strengths existed. The information signals to those working with the student deeper examination in terms of the student's competencies and skills in the social, behavioral, and emotional domains.

Applying partial eta squared, parent and teacher rater informants showed meaningfully significant differences on the Family Involvement and Intrapersonal subscales. Items in these subscales included "demonstrates a sense of belonging to family," and "enjoys a hobby" (Epstein & Pierce, in press). Each of these informants viewed the student in different environmental settings and might not have had a full picture of the attributes of the student. Harkening back to the underlying belief regarding the impact of the environmental context on behavior, a limitation of using these results derived from CTT was behavior changed according to environmental variables. When making determinations about the use of a multi-rater measurement, it is essential that those doing the rating are familiar with the measure and its constructs and are familiar with the student in relation to those constructs.

At the beginning of this study, I hypothesized there would be variations between ratings based on informant type at a magnitude that would further inform choice of raters as well as provide critical information to incorporate into administrative training sessions for the use of strengths-based behavioral rating scales. Before applying g study analysis to this data set, the results obtained led me to believe that even though student raters showed large variations in data, generally there was no difference between parent and teacher raters in this study design. However, applying G theory analysis to the data painted a significantly different picture. When evaluating information from all three rater types together, the student rater significantly impacted the overall understanding of the needs of the students and therefore could impact service implementation and goal development for the student.

Relative Generalizability Coefficient Variance

The relative *g* coefficient is applicable for “quantifying degrees of consistency of the relative standings of individuals, but not the consistency of actual scores” (Fan & Sun, 2014, p. 14). It is used for norm-referenced assessment decisions such as progress monitoring. In this study, relative variance components, *person x rater* (50.1%) and *person x item* (26.8%), had the highest levels of possible variance in the universe of observations. *Person x rater x item* was 23% (see Figure 3). Applying the relative *g* coefficient, *student x rater* type, yielded the largest variance, indicating disagreement among rater informants regarding perceptions of students. A large variance for person x rater interaction suggested individual raters had inconsistent perceptions of certain students (e.g., the rater type influenced the scores of the student). Multiple informants might rate students differently based on their relationship with the student, the characteristics of a class or family unit (e.g., group or family dynamics), or even their interpretation of BERS statements. Rater, item, and the rater and item interaction held no percentage of relative error variance.

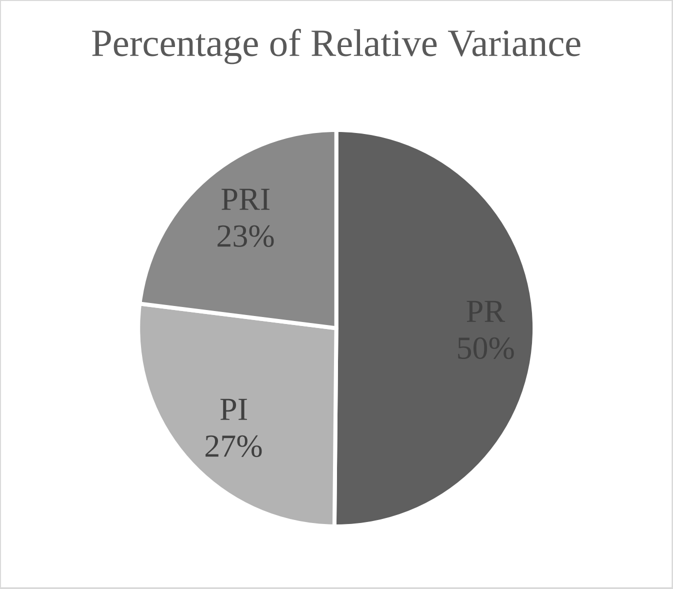


Figure 3. Percentage of variance explained—relative variance.

A 26.8% relative variance attributed to the *person x item* interaction could mean inconsistencies between subscales for a student, averaging over the type of raters. There was variance due to the behavior being evaluated (e.g., internalizing behavior). In other words, this told us the scores given to students differed in relation to different items. This could indicate specific construct variance existed. Construct variance would be of greater concern in absolute decisions.

Specific to relative decisions using the BERS-3 (Epstein & Pierce, in press), variance could be the result of using multiple items to evaluate each person and aggregating scores across items. If the length of the BERS was increased by a certain percentage of items, internal consistency reliability might improve. In terms of rater training, this information suggested more specific training tailored to a deeper understanding of the construct of interest being measured s necessary. It is important that all raters have the same understanding of the underlying idea of the construct under investigation. The remaining 23% of relative variance was attributable to the *person* x *rater* x *item* source of variance. This meant the *person* x *rater* interactions were inconsistent across items within the subscale. To mitigate this constraint in the future in order to determine the source of the variance from *person* x *rater* x *item,* adding a third facet, i.e., observation, would help pinpoint what was a temporary variance of error and what was a permanent variance of error (Lakes & Hoyt, 2009). When educators use the BERS-3 for relative decision making, it would be best to collect data on more than one occasion from all three rater types. Doing so would increase the reliability of the data. Outcomes of the data derived during this study indicated a single administration of the assessment would not provide accurate and adequate data for decision making.

Generalizability theory analysis or g facets analysis is an output the researcher could run after the data set was entered into EduG v. 6. Running this analysis allowed me to evaluate a universe of errors by looking more closely at inconsistencies among rater types. The *g* coefficient showed a reliability estimate that was too low for accurate information gathering (.54). Borrowing reliability coefficient standards from CTT, the minimum acceptable level was .80. Decisions regarding screening for social, emotional, and behavioral supports and services for students require an even higher criterion to be employed of .90 or even .95 (Schumacker, 2010). In the g study, the *g* coefficient (.54) not only showed a high level of variance between rater type, it also showed almost a 40% chance of error using the parameters of this study. This would significantly influence decisions based on data gathered. Again, knowing where the error originated was not available using CTT. The type of rater completing the measure produced an unacceptable error level in the ratings of students on this strengths-based behavioral rating scale and must be considered for planning and procedural components of school-based assessment.

Table 21 presents a g facets analysis for the *p* x *r* x *i* g study design. Looking specifically at the relative *g* coefficient, the rater type with the highest reliability was the parent rater (0.6), followed by the teacher rater type (0.5), with the student rater type being the least reliable (0.1). This told me that less variation occurred when parents used the BERS-3 Epstein & Pierce, in press) to rate their child and the most accurate information was gleaned from this type of rater. In terms of testing protocols, the student self-report might provide little useable information for these types of decisions.

Table 21

Generalizability Facets Analysis

Facet	Level	Relative G Coefficent	Absolute G Coefficient
Rater	Student	0.1	0.0
	Parent	0.6	0.1
	Teacher	0.5	0.1
Item	Interpersonal Strength	0.5	0.1
	Family Involvement	0.5	0.1
	Intrapersonal Strength	0.5	0.1
	School Functioning	0.5	0.1
	Affective Strength	0.6	0.1
	Strengths Index Composite	0.6	0.4

The relative *g* coefficient for constructs measured in the Affective Strengths and Strengths Index subscales was the highest at 0.6, followed by the Interpersonal Strength, Family Involvement, Intrapersonal Strength, and School Functioning subscales, all with a score of 0.5. This showed that when used together (e.g. all subscales working toward one overall picture), the BERS-3 (Epstein & Pierce, in press) had a greater reliability than the sum of its parts.

Absolute Generalizability Coefficient

Variance

The absolute *g* coefficient is an indicator of reliability for absolute decision making such as high-stakes decisions or diagnostic purposes. Parsing absolute sources of variance for this data, 0.5% was from rater, 90.4% was from the representative items of each construct, 4.5% was from the *person* x *rater* interaction, 2.4% was from the *person* x *item* interaction, 0.1% was from the *rater* x *item* interaction, and a total of 2.1% was from the *person* x *rater* x *item* interaction variance source (see Figure 4). The 0.5% variance among raters indicated systematic differences in raters due to bias. Variance in the item facet (90.4%) suggested consistently higher ratings for several students on some subscales and not others. The interaction variation of 4.45% for *person* x *rater* implied individual raters had inconsistent perceptions of certain students. The *person* x *item* interaction pointed to inconsistences between subscales for a student when averaged over rater type. The *person* x *rater* x *item* source of variance was very small, signifying scant inconsistencies across items within the subscale. This further demonstrated the behavior the subscale purported to measure seemed to be reliable. However, for absolute decision making, rater bias must be considered. Collecting multiple sources of data remains important and necessary in decisions regarding special education and related services for students with ED being considered for services under IDEA (2004) or who are currently served under IDEA.

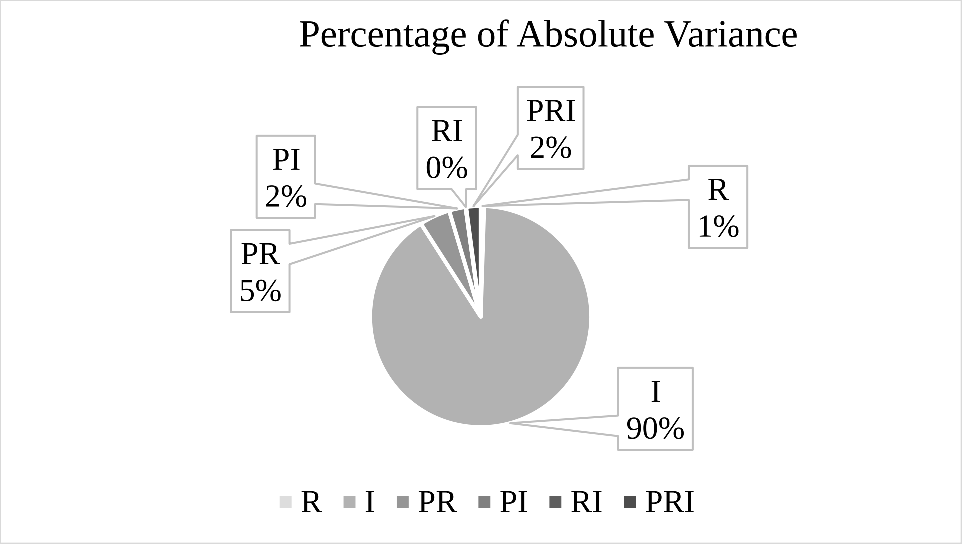


Figure 4. Percent of variance explained—absolute variance.

It is important to remember the BERS-3 (Epstein & Pierce, in press) is a norm-referenced assessment used primarily for progress monitoring and service decision purposes in schools and clinics. In tandem with a relative *g* coefficient, an absolute *g* coefficient is produced from conducting a G theory analysis. In this study, the absolute *g* coefficient was .10. This made sense due to the nature of the assessment tool as it is a norm-referenced measurement. Almost all the variance came from the subscale or item facet of generalization. *Rater*, *person* x *rater* interaction, and *person* x *item* interaction had small levels of variance—all around 2-4% of the total universe of possible variance for the absolute analysis. The item percentage let us know when administering the BERS-3 for data collection purposes leading to diagnostic decision making, it should be paired with further evaluation, observations, teacher reports, as well as other forms of collected data to begin to mitigate this finding. One interesting data outcome to note was the Strengths Index subscale showed better reliability for high-stakes decision making than all other subscales.

Additional Generalizability and Decision Studies

Generalizability Studies Examining Gender

Relative variance: g x r x i. It is not unusual in g study research to continue and expand the g study after conducting the initial g study design. After the original design research was complete, the variable of gender and its potential impact on the relative error variance and subsequent reliability of the measure was examined. I hypothesized that gender would factor into rater biases among the three rater types. First, I examined the facet gender (see Table 22--Generalizability Study Observation and Estimation Designs for *g* x *r* x *i*, Table 23--Analysis of Variance using EduG v. 6 for the *g* x *r* x *i* study design, and Table 24-- Generalizability Study Analysis using EduG v. 6 for the *g* x *r* x *i* study design).

Table 22

Generalizability Study Observation and Estimation Designs for g x r x i

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
Gender	G	2	INF	
Rater	R	3	INF	
Item-Subscale	I	6	INF	

Table 23

Analysis of Variance using EduG v. 6 for the g x r x i Study Design

Source	SS	df	MS	Components				
				Random	Mixed	Corrected	%	SE
G	110.9	1	110.9	-0.5	-0.5	-0.5	0.0	6.7
R	18.2	2	9.1	-8.0	-8.0	-8.0	0.0	6.7
I	36188.0	5	7237.6	1203.6	1203.6	1203.6	97.2	644.6
GR	224.7	2	112.3	15.9	15.9	15.9	1.3	13.3
GI	118.8	5	23.8	2.2	2.2	2.2	0.2	4.8
RI	92.8	10	9.3	-3.9	-3.9	-3.9	0.0	4.0
GRI	170.3	10	17.0	17.0	17.0	74.9	1.4	7.0
Total	36923.7	35					100.0	

Results from the secondary g study (*g* x *i*) indicated a smaller relative variance error interaction percentage (5.6%). Through the lens of gender, the construct or item variance was not as significant since there were fewer inconsistencies from one item to another according to gender, averaging over raters. The relative error of variance percentage for *gender* x *rater* interaction was 80.1%. There was disagreement among rater types that could be specifically attributed to the gender facet. Multiple informants rated individuals differently based on gender alone. This could be attributable to biases and stereotypes raters held regarding gender. When planning and implementing training for raters, these results indicated training specific to gender biases was needed.

Table 24

Generalizability Study Analysis using EduG v. 6 for the g x r x i Study Design

Source of variance	Differentiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
G	(0.0)		.....		.....	
	.....	R	.....		(0.0)	0.0
	.....	I	.....		200.6	96.8
	.....	PR	5.3	80.1	5.3	2.6
	.....	PI	0.4	5.6	0.4	0.2
	.....	RI	.....		(0.0)	0.0
	.....	PRI	0.9	14.3	0.9	0.5
Sum of variances	0.0		0.0	100.0	207.2	100.0
SD	0.0		Relative SE : 2.6		Absolute SE :14.4	
Relative Coefficient *G*	0.00					
Absolute Coefficient *G*	0.00					

Grand mean for levels used: 35.2
Variance error of the mean for levels used: 203.9
Standard error of the grand mean: 14.3

In the original study design, the relative error variance for *p* x r x *i* interaction was 23%, indicating inconsistencies across items within the subscales. The follow-up g study examining gender reduced this percentage to 14.3%. Examining variance more specifically through gender allowed me to understand variance in the relative scoring of the BERS-3 (Epstein & Pierce, in press). Controlling for gender, the behaviors the subscales purported to measure were more reliable than when examining this without controlling for gender. This might indicate more training for all rater types was needed to ensure reliable assessment of the construct under examination. Administrative protocols to reduce gender biases are warranted. Results indicated a better awareness of the role gender plays on how raters measured behavioral constructs was needed.

Results from the secondary g study (*g* x *i*) indicated a smaller relative variance error interaction percentage (5.6%). Through the lens of gender, the construct or item variance was not as significant since there were fewer inconsistencies from one item to another according to gender, averaging over raters. The relative error of variance percentage for *gender* x *rater* interaction was 80.1%. There was disagreement among rater types that could be specifically attributed to the gender facet. Multiple informants rated individuals differently based on gender alone. This could be attributable to biases and stereotypes raters held regarding gender. When planning and implementing training for raters, these results indicated training specific to gender biases was needed.

In the original study design, the relative error variance for *p* x r x *i* interaction was 23%, indicating inconsistencies across items within the subscales. The follow-up g study examining gender reduced this percentage to 14.3%. Examining variance more specifically through gender allowed me to understand variance in the relative scoring of the BERS-3 (Epstein & Pierce, in press). Controlling for gender, the behaviors the subscales purported to measure were more reliable than when examining this without controlling for gender. This might indicate more training for all rater types was needed to ensure reliable assessment of the construct under examination. Administrative protocols to reduce gender biases are warranted. Results indicated a better awareness of the role that gender plays on how raters measure behavioral constructs was needed.

Absolute variance: g x r x i. It is important to note the BERS-3 (Epstein & Pierce, in press), commonly used for the evaluation of pre-referral services for placement in specialized services, measuring outcomes of services, and the identification of individual behavioral and emotional strengths along with places for strengths development, is not a diagnostic tool. However, as a tool for the identification of strengths along with areas for growth, I felt it remained important to examine the facet gender in terms of absolute decision making. Absolute error variance for the factor, item, accounted for 96.8% of all variance in the data, suggesting much higher ratings for several students on some subscales and not others. The *gender* x *rater* interaction accounted for 2.6% of absolute error variance and the *gender* x *rater* x *item* interaction accounted for 0.5% of absolute error variance. The *gender* x *rater* interaction showed individual raters had inconsistent perceptions based on the gender of the student. The 0.5% absolute error variance from the *gender* x *rater* x *item* interaction signaled inconsistencies arising among items on subscales because of the gender facet.

Generalizability facet analysis*: g x r x i.* Similar to the original study, gender as a facet was examined by running a g facet analysis using EduG v.6. Results from the g facet analysis using gender as a facet are embedded within the text of this section. The relative *g* coefficient for *teacher* x *gender* proved to be the most reliable with a score of .70. Borrowing reliability coefficient standards from CTT, the minimum acceptable level is .80. However, the g facet analysis, broken down into gender and rater type, presented acceptable reliability between gender and the teacher type rater. From this, one might infer teacher raters rating the same student on the BERS-3 (Epstein & Pierce, in press) as a parent and the student themselves were less influenced by the gender of the student under examination. In terms of training for multiple informants, this could indicate parents might need more coaching and understanding not only on what was appropriate for their child at that moment in their development and what was appropriate for their child at that moment in their development as it related specifically to gender.

Generalizability facet analysis for the item (subscale) facet through the secondary g study using *gender* x *rater* x *item* was 34.1% for the School Functioning subscale. When looking at gender and rater, the biggest variation was between female self-reports on the Interpersonal Strengths subscale who rated themselves very high (*M* = 39.8) and male self-reports who rated themselves generally very low (*M* = 30.3). This meant the male students rated themselves with a higher probability of having ED and the girls rated themselves with a lower probability of having ED. Gender and rater interaction had a significant impact on subscale scores. Looking at *gender* x *item* interaction facets, the female student reports for School Functioning (41.2%) and male student reports for the Strengths Index subscale (49.9%) had the most variation (41.2%) from the mean. The student *rater type* x *subscale* (item) experienced the most error variance (81%). This was an indication that when optimizing measurement conditions for students, gender differences should be considered.

Additional Generalizability Study Examining Two Rater Types Only: Parent and Teacher

Relative variance: Parent and teacher rater types. The results from the ANOVA using SPSS v. 25 indicated possible positive illusory bias among students; this finding led me to conduct a second alternate g study. For this design, I took the student rater forms out of the database. I was left with 23 completed BERS-3(Epstein & Pierce, in press) forms by two rater types, parents and teachers, across all six subscales. This gave me 276 data points. For the relative error variance, the *perso*n x *rater* interaction accounted for 50.4%, the *person* x *item* interaction had a 28.8% error variance, and the *person* x *rater* x *item* interaction had a 20.9% error variance. These percentages were close to the outcomes from the original g study examining three rater types: (a) parent, (b) teacher, and (c) student. It is important to point out the *person* x *rater* x *item* interaction was lower by 2.1% for this g study design. While not a huge difference, this source of variance was significant because it showed that controlling for the student rater facet, inconsistencies across items within the subscale were reduced and the behavior the subscale purported to measure increased in reliability. The relative *g* coefficient for both rater types were 50%. Finally, like the original study, this g study gave indications that to create the most optimal measurement conditions, rater biases need to be taken into account.

Absolute variance: Parent and teacher rater types. For absolute error variance, the rater facet alone accounted for 0.5% variance, item facet accounted for 76.5% variance, *person* x *rater* interaction accounted for 11.5% variance, *person* x *item* interaction accounted for 6.6% variance, *rater* x *item* interaction accounted for.1% variance, and the *person* x *rater* x *item* interaction error variance was 4.8%. Comparing this g study design to the original, the item error variance (90.4%) was reduced significantly when controlling for rater types (76.5%). Additionally, the *person* x *rater* interaction (4.5% originally), *person* x *item* interaction (2.4% originally), and the *person* x *rater* x *item* interaction (2.1% originally) were all significantly reduced. There were fewer extreme scores on some subscales, there were more consistent perceptions of students, and the inconsistencies between subscales for a student when averaged over rater type were less than when accounting for the student rater type in addition to the parent and teacher raters. Inconsistencies across items within the subscale were negligible, indicating the behavior the subscale purported to measure seemed to be reliable. By adjusting rater type to exclude the student self-report, the accuracy of scores increased. This was significant as the results would be used to make decisions regarding special education and related services. The absolute *g* coefficient was 0.1 for parent raters and 0.2 for teacher raters; both were too low for absolute decision making. This was not alarming as the BERS-3 was not intended for absolute decisions and diagnostic decisions.

Design Studies

In addition to the alternate g study, a brief d study followed the original research analysis (*p* x *r* x *i*). For the d study analysis, the universe of possible error for rater type was controlled by changing the random facet of three raters to 4, 5, 6, 7, and 8 rater levels (see Table 25). While the absolute *g* coefficient did not improve, the relative *g* coefficient rose to 0.69. For the type of decisions being made from the data, this level remained too low with a preferable reliability coefficient of .90 or .95 (Schumacker, 2010). Adding an eighth rater increased the relative *g* coefficient significantly from 0.54 to 0.69 but there was no difference with the results for the absolute *g* coefficient. Both absolute and relative error variances were reduced but not at the level needed for optimal measurement conditions and subsequent data derived from those measurements. The BERS-3 (Epstein & Pierce, in press) was used for relative decision making, i.e., for progress monitoring and service implementation decisions, rather than for diagnostic determinations. Gathering information on the student's levels of behavioral strengths using the BERS-3 would be best done when a wide range of raters were asked to provide data and multiple data points were used. As the number of raters increases so do the levels of reliability.

A limitation in this d study was the inclusion of the student rater type. Previously, when examining the g facet analysis, student self-report variance was statistically significant and thus impacted the overall reliability ratings for the BERS-3 (Epstein & Pierce, in press). Potentially due to the phenomenon of positive illusory bias, the data derived from the BERS-3 student rater form created an inaccurate picture of student need, competencies, and strengths. However, taking only the parent and teacher forms into account for relative decisions like IEP service implementation and selection and IEP goals, the information gleaned from the BERS-3 was invaluable. Even though caution needs to be applied when decisions are derived from data collected via the student self-report, the data could be useful when working with the student to build goals and to leverage perceived competencies and skills.

Table 25

Design Study Optimization

				Opt.	1	Opt.	2	Opt.	3	Opt.	4	Opt.	5
		Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
	P	23	INF	23	INF	23	INF	23	INF	23	INF	23	INF
	R	3	INF	4	INF	3	INF	4	INF	5	INF	6	INF
	I	6	INF	6	INF	6	INF	6	INF	6	INF	6	INF
Observations		414		552		690		828		966		1104	
Coef_*G* rel.		0.543		0.593		0.627		0.652		0.671		0.686	
rounded		0.54		0.59		0.63		0.65		0.67		0.69	
Coeff_ *G* abs.		0.09		0.097		0.098		0.099		0.099		0.100	
Rounded		0.10		0.10		0.10		0.10		0.10		0.10	
Rel. Err. Var.		18.1		14.8		12.8		11.5		10.5		9.8	
Rel. Std. Err.M.		4.3		3.8		3.6		3.4		3.2		3.1	
Abs. Err. Var.		202.9		199.3		197.1		195.6		194.6		193.8	
Abs. Std. Err.M		14.2		14.1		14.0		14.0		13.9		13.9	

Note: Univ. represents the universe of error variance; Lev. represents the level or number of raters, items, and people associated with each facet. INF indicates that the facet is random and part of an infinite universe.

After the d study analysis was run for the original g study (all three rater types), the parent and teacher informants were examined to find out more about measurement condition optimization. To provide the most reliable data, the universe of observation for the facets of rater type and subscale had to be fixed rather than random. In the universe of possible observation of two raters, parent and teacher, and only six subscales, both the relative and absolute g coefficients (1.0) indicated the highest levels of reliability possible. Additionally, while not perfect, in the second optimization d study, the relative error variance was reduced significantly when the universe of observation consisted of two rater types and the subscale universe of observation was random and infinite. This meant that when looking for an optimal design for measurement conditions, the two raters, parent and teacher, sufficed and the self-report was not needed. Appropriate and accurate information was gathered from the parent and teacher forms of the BERS-3 (Epstein & Pierce, in press).

Conclusion

The current study explored variances using a strengths-based assessment approach, an approach known to ease associated stigma, leverage a student’s competencies, and bolster parent and school partnerships and communication (Epstein et al., 2000). It was hypothesized that rater-type would explain the largest amount of variance on the strengths-based behavioral rating scale—BERS-3 (Epstein & Pierce, in press). Millman (2015) found over 20% variance in between-teacher at-risk assessment scores and Johnson et al. (2016) used the Direct Behavior Rating-Single Item Scale to measure teacher and classroom variance using G theory with a result of 20% to 20.6% variance between teacher types. However, in the current study, rater facet alone did not indicate error variance. It was the interaction of *person* x *rater*, *person* x *item*, and *person* x *rater* x *item* that produced the greatest amounts of error. An area for further research would include gathering factors that explained the rationale for this outcome, e.g., whether it resulted from training procedures and method applied, the place the raters filled out the report, the time of day the rating occurred, and rater biases or non-biases of students. All of these factors or facets in G theory could be analyzed through the lens of G theory to understand more fully the impact on reliability and are recommended for future studies.

The person variance was expected to be high and all other sources of variance low; however, this was not the result of this study. Interaction effects of *person* x *rater*, *person* x *item*, and *person* x *rater* x *item* were the only areas in which error variance occurred. The *person* x *item* variance of 56% could be understood as there would naturally be differences between the individual student's externalizing and internalizing behaviors and the item type. The interaction between the *person* x *rater* might indicate no overall error variance due to the rater type but rather differences among raters more specific to the individual student. This could still be a problem since there might have been student characteristics that caused some inconsistences in ratings. Rather than the average rating generated by all the raters who could have possibly conducted the rating, instead it was on the *person* x *rater* level where variance occurred. Adding student characteristics as facets to investigate why informant ratings differed dependent upon the student was important. In this study, the facet, gender, was added. Future studies, examining disability status, socio-economic status, and racial or ethnic variables might provide more information on potential rater biases.

Although the specific combinations of items and types of raters identified in the current study must be considered specific to the BERS-3 (Epstein & Pierce, in press), these results have broader implications for the use of strengths-based behavioral rating scales for data-based decision making. Unfortunately, G theory has not been fully adopted within the literature to date given the seriousness of collecting appropriate and adequate data to provide access to FAPE. Rather than examining psychometric assessments through a narrow lens (CTT), which does not allow for a greater understanding of the factors affecting the errors within scores, applying G theory allows the identification of the most efficient assessment design by testing for the universe of errors due to single factors, such as the classroom environment or the time of day, as well as the interaction of a combination of facets, such as the occasion of measurement and the rater. As hypothesized above, if certain student characteristics were at least partly responsible for some of the raters' inconsistencies in scoring, this would be an important factor to consider when developing training protocols for raters. The most efficient way to administer a behavior rating scale might not be the most acceptable or sustainable and by applying G theory analysis, the informant has the flexibility of taking acceptability into account without sacrificing the accuracy and reliability of the resultant data for school-based decision making.

REFERENCES

Achenbach, T. M. (2014). Achenbach system of empirically based assessment (ASEBA). In *The encyclopedia of clinical psychology*, (pp. 31-29). New York: Springer.

Algina, J., & Swaminatahan, H. (2015). Psychometrics: Classical test theory. In J. D. Wright (Ed.), *International encyclopedia of social and behavioral sciences* (pp. 423-430). New York, NY: Elsevier.

Allen, A. N., Kilgus, S. P., Burns, M. K., & Hodgson, C. (2018). Surveillance of internalizing behaviors: A reliability and validity generalization study of universal screening evidence. *School Mental Health*, *11*(2), 194-209.

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

Becker, S. P., Paternite, C. E., Evans, S. W., Andrews, C., Christensen, O. A., Kraan, E. M., & Weist, M. D. (2011). Eligibility, assessment, and educational placement issues for students classified with emotional disturbance: Federal and state-level analyses. *School Mental Health*, *3*(1), 24-34.

Bergeron, R., Floyd, R. G., McCormack, A. C., & Farmer, W. L. (2008). The generalizability of externalizing behavior composites and subscale scores across time, rater, and instrument. *School Psychology Review*, *37*(1), 91–108.

Bloch, R., & Norman, G. (2012). Generalizability theory for the perplexed: A practical introduction and guide: *Medical Teacher, 34*(11), 960-992.

Bracken, B. A., & Keith, L. K. (2004*). Clinical assessment of behavior*. Lutz, FL: Psychological Assessment Resources.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27-34.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*(4), 295-317.

Brennan, R. L. (2003). *Coefficients and indices in generalizability theory*. Retrieved from https://education.uiowa.edu/sites/education.uiowa.edu/files/documents/centers/casma/publications/casma-research-report-1.pdf

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

Briesch, A. M., Chafouleas, S. M., & Johnson, A. (2016). Use of generalizability theory within K-12 school-based assessment: A critical review and analysis of the empirical literature. *Applied Measurement in Education, 29*(2), 83-107.

Briesch, A. M., Chafouleas, S. M., & Riley-Tillman, T. C. (2010). Generalizability and dependability of behavioral assessment methods: A comparison of systematic direct observation and direct behavior rating. *School Psychology Review*, *39*(3), 408–421.

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology, 52*(1), 13-35.

Bruckner, C. T., Yoder, P. J., & McWilliam, R. A. (2006). Generalizability and decision studies: An example using conversational language samples. *Journal of Early Intervention, 28*(2), 139–153.

Bruhn, A. L., Woods-Groves, S., & Huddle, S. (2014). A preliminary investigation of emotional and behavioral screening practices in K-12 schools. *Education & Treatment of Children, 37*(4), 611-634.

Buckley, J. A., & Epstein, M. H. (2004). The Behavioral and Emotional Rating Scale-2 (BERS-2): Providing a comprehensive approach to strengths-based assessment. *The California School Psychologist*, *9*(1), 21-27.

Chafouleas, S. M., Briesch, A. M., Riley-Tillman, T. C., Christ, T. J., Black, A. C., & Kilgus, S. P. (2010). An investigation of the generalizability and dependability of Direct Behavior Rating Single Item Scales (DBR-SIS) to measure academic engagement and disruptive behavior of middle school students. *Journal of School Psychology*, *48*(3), 219-246.

Chafouleas, S. M., Christ, T. J., Riley-Tillman, T., Briesch, A. M., & Chanese, J. A. M. (2007). Generalizability and dependability of direct behavior ratings to assess social behavior of preschoolers. *School Psychology Review, 36*(1), 63-79.

Chafouleas, S. M., Riley-Tillman, T. C., & Christ, T. J. (2009). Direct behavior rating (DBR): An emerging method for assessing social behavior within a tiered intervention system. *Assessment for Effective Intervention, 34*(4), 195–200.

Christ, T. J., Riley-Tillman, T. C., & Chafouleas, S. M. (2009). Foundation for the development and use of direct behavior rating (DBR) to assess and evaluate student behavior. *Assessment for Effective Intervention, 34*(4), 201-213.

Christ, T. J., Riley-Tillman, T. C., Chafouleas, S. M., & Boice, C. H. (2010). Direct Behavior Rating (DBR): Generalizability and dependability across raters and observations. *Educational and Psychological Measurement, 70*(5), 825–843.

Climie, E., & Henley, L. (2016). A renewed focus on strengths-based assessment in schools: Strengths-based assessment. *British Journal of Special Education, 43*(2), 108-121.

Conger, A. J., Conger, J. C., Wallander, J., Ward, D., & Dygdon, J. (1983). A generalizability study of the Conners' teacher rating scale-revised. *Educational and Psychological Measurement, 43*(4), 1019-1031.

Conners, C. K. (1969). A teacher rating scale for use in drug studies with children. *American Journal of Psychiatry, 126*(6), 884-888.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis*. Upper Saddle River, NJ: Pearson Education, Inc.

Coulacoglou, C., & Saklofske, D. H. (2017). *Psychometrics and psychological assessment: Principles and applications*. London, United Kingdom: Academic Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York, NY: John Wiley.

Crowley, S. L., Thompson, B., & Worchel, F. (1994). Validity studies the children's depression inventory: A comparison of generalizability and classical test theory analyses. *Educational and Psychological Measurement, 54*(3), 705–713.

Cullinan, D., & Epstein, M. H. (2013a). Development, reliability, and construct validity of the emotional and behavioral screener. *Preventing School Failure: Alternative Education for Children and Youth, 57*(4), 223-230.

Cullinan, D., & Epstein, M. H. (2013b). *Emotional and behavioral screener*. Austin, TX: PRO-ED.

Cullinan, D., Harniss, M., Epstein, M. H., & Ryser, G. (2001). The scale for assessing emotional disturbance: concurrent validity. *Journal of Child and Family Studies*, *10*(4), 449-466.

Dowdy, E., & Kim, E. (2012). Choosing informants when conducting a universal screening for behavioral and emotional risk. *School Psychology Forum, 6*(4), 98-107.

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and Perspectives, 38*(1), 105-124.

Duppong Hurley, K., Lambert, M. C., Epstein, M. H., & Stevens, A. (2015). Convergent validity of the strengths-based behavioral and emotional rating scale with youth in a residential setting. *The Journal of Behavioral Health Services & Research, 42*(3), 346-354.

Ebel, R. L. (1972). *Essentials of educational measurement* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Epstein, M. H. (2000). The behavioral and emotional rating scale: A strengths-based approach to assessment. *Diagnostique*, *25*(3), 249–256.

Epstein, M. H. (2004). *Behavioral and emotional rating scale: A strengths-based approach to assessment: Examiner's manual*. Austin, TX: Pro-Ed.

Epstein, M. H., Cullinan, D., Pierce, C. D., Huscroft-D'Angelo, J., & Wery, J. J. (in press). *Scales for Assessing Emotional Disturbance* (3rd ed.)*: Examiner's manual.* PRO-ED.

Epstein, M. H., & Pierce, C. D. (in press). *Behavioral and Emotional Rating Scale – Examiner's manual* (3rd ed.). PRO-ED.

Epstein, M. H., Rudolph, S., & Epstein, A. A. (2000). Using strengths-based assessment in transition planning. *Teaching Exceptional Children, 32*(6), 50.

Epstein, M., & Sharma, J. M. (1998). *Emotional and behavioral rating scale: A strengths based approach to assessment.* Austin, TX: Pro-Ed.

Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015).

Fan, C. H., & Hansmann, P. R. (2015). Applying generalizability theory for making quantitative RTI progress-monitoring decisions. *Assessment for Effective Intervention*, *40*(4), 205-215.

Fan, X., & Sun, S. (2014). Generalizability theory as a unifying framework of measurement reliability in adolescent research. *The Journal of Early Adolescence, 34*(1), 38-65.

Frey, B. (2018a). Classical test theory. In *The SAGE encyclopedia of educational research, measurement, and evaluation* (Vols. 1-4, pp. 278-283). Thousand Oaks, CA: SAGE Publications, Inc. doi:10.4135/9781506326139

Frey, B. (2018b). Rasch model. In *The SAGE encyclopedia of educational research, measurement, and evaluation* (Vols. 1-4, pp. 1370-1373). Thousand Oaks, CA: SAGE Publications, Inc.

Gage, N. A., Adamson, R., Mitchell, B. S., Lierheimer, K., O'Connor, K. V., Bailey, N., … Jones, S. (2010). Promise and possibility in special education services for students with emotional or behavioral disorders: Peacock Hill revisited. *Behavioral Disorders, 35*(4), 294-307.

Gage, N. A., & Lierheimer, K. (2012). Exploring self-concept for students with emotional and/or behavioral disorders as they transition from elementary to middle school and high school. *Education Research International, 2012,* 1-11.

Gage, N. A., Prykanowski, D., & Hirn, R. (2014). Increasing reliability of direct observation measurement approaches in emotional and/or behavioral disorders research using generalizability theory. *Behavioral Disorders, 39*(4), 228-244.

Goldring, E. B., Berends, M., & American Association of School Administrators. (2009). *Leading with data: Pathways to improve your school*. Thousand Oaks, CA: Corwin Press.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*(5), 581-586.

Gresham, F. M., Dart, E. H., & Collins, T. A. (2017). Generalizability of multiple measures of treatment integrity: Comparisons among direct observation, permanent products, and self-report. *School Psychology Review, 46*(1), 108-121.

Hanchon, T. A., & Allen, R. A. (2018). The identification of students with emotional disturbance: Moving the field toward responsible assessment practices. *Psychology in the Schools, 55*(2), 176-189.

Hauenstein, C. E., & Embretson, S. E. (2019). Development and application of item response theory. In J. E. Edlund & and A. L. Nichols (Eds.), *Advanced research methods for the social and behavioral sciences* (pp. 309-327). New York: Cambridge University Press.

Hays, R. D., Brown, J., Brown, L. U., Spritzer, K. L., & Crall, J. J. (2006). Classical test theory and item response theory analyses of multi-item scales assessing parents' perceptions of their children's dental care. *Medical Care, 44*(11), S60-S68.

Hintze, J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review, 33*(2),258–270.

Hunsley, J., & Mash, E. J. (2018). *A guide to assessments that work* (2nd ed.). New York: Oxford University Press.

Individuals With Disabilities Education Act, 20 U.S.C. § 1400 (2004).

Ikeda, M. J., Neessen, E., & Witt, J. C. (2008). Best practices in universal screening. In A. Thomas, A. & Grimes, J. (Eds.), *Best practices in school psychology* (5th ed., pp. 103-114). Bethesda, MD: National Association of School Psychologists.

Johnson, A. H., Miller, F. G., Chafouleas, S. M., Welsh, M. E., Riley-Tillman, C., & Fabiano, G. (2016). Evaluating the technical adequacy of DBR-SIS in tri-annual behavioral screening: A multisite investigation. *Journal of School Psychology, 54*, 39–57.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6,* 125–160.

Kauffman, J. M., & Landrum, T. J. (2018). *Characteristics of emotional and behavioral disorders of children and youth with disabilities* (11th ed.). Upper Saddle River, NJ: Pearson.

Kaurin, A., Egloff, B., Stringaris, A., & Wessa, M. (2016). Only complementary voices tell the truth: A reevaluation of validity in multi-informant approaches of child and adolescent clinical assessments. *Journal of Neural Transmission*, *123*(8), 981-990.

Kilgus, S. P., & von der Embse, N. P. (2014). *Social, academic, and emotional behavior risk screener* (Unpublished technical manual).

Kovacs, M., & Beck, A. T. (1977). An empirical-clinical approach toward a definition of childhood depression. In J. G. Schulterbrandt & A. Raskin (Eds.), *Depression in childhood: Diagnosis, treatment, and conceptual models* (pp. 1–25). New York: Raven Press.

Kratochwill, T. R., Clements, M. A., & Kalymon, K. M. (2007). Response to intervention: Conceptual and methodological issues in implementation. In *Handbook of response to intervention* (pp. 25-52). Boston, MA: Springer.

Lakes, K. D., & Hoyt, W. T. (2009). Applications of generalizability theory to clinical child and adolescent psychology research. *Journal of Clinical Child and Adolescent Psychology*, *38*(1), 144–165.

Lambert, M. C., January, S. A., Epstein, M. H., Spooner, M., Gebreselassie, T., & Stephens, R. L. (2015). Convergent validity of the behavioral and emotional rating scale for youth in community mental health settings. *Journal of Child and Family Studies, 24*(12), 3827-3832.

Lloyd, B. P., Bruhn, A. L., Sutherland, K. S., & Bradshaw, C. P. (2019). Progress and priorities in research to improve outcomes for students with or at risk for emotional and behavioral disorders. *Behavioral Disorders*, *44*(2), 85–96.

Lomax, R. G. (1982). An application of generalizability theory to observational research. *The Journal of Experimental Education*, *51*(1), 22–30. doi:10.1080/00220973.1982.11011835

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. New York: Routledge.

McWilliam, R. A., & Ware, W. B. (1994). The reliability of observations of young children’s engagement: An application of generalizability theory. *Journal of Early Intervention*, *18*(1), 34–47.

Mason, B. A., Gunersel, A. B., & Ney, E. A. (2014). Cultural and ethnic bias in teacher ratings of behavior: aAcriterion-focused review. *Psychology in the Schools*, *51*(10), 1017-1030.

Mathur, S., & Jolivette, K. (2012). Chapter 4: Placement of students with emotional and behavioral disorders. In J. Bakken, F. Obiakor, & A. Rotatori (Eds.), *Behavioral disorders: Identification, assessment, and instruction of students with EBD* (Vol. 22, pp. 87-105). London: Emerald Group Publishing Limited.

Merikangas, K. R., He, J., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., … Swendsen, J. (2010). Lifetime prevalence of mental disorders in US adolescents: Results from the National Comorbidity Survey Replication-Adolescent Supplement (NCS-A). *Journal of the American Academy of Child and Adolescent Psychiatry, 49*(10), 980–989.

Miller, F. G., Crovello, N. J., & Chafouleas, S. M. (2017). Progress monitoring the effects of daily report cards across elementary and secondary settings using direct behavior rating: Single Item Scales. *Assessment for Effective Intervention*, *43*(1), 34-47.

Millman, M. K. (2015). *Differences between-teacher-reports on universal risk assessments: Exploring the teacher's role in universal screening of student behavior* (Doctoral dissertation). Miami University, FL.

Mitchell, B. S., Kern, L., & Conroy, M. A. (2019). Supporting students with emotional or behavioral disorders: State of the field. *Behavioral Disorders*, *44*(2), 70–84.

Morgan, D. L. O. (2001). *Estimating the teacher /classroom impact on individual score variance: A combination of generalizability theory and two-level modeling* (Order ProQuest Dissertations & Theses Global #276217814).

Mooney, P., Epstein, M. H., Ryser, G., & Pierce, C. D. (2005). Reliability and validity of behavioral and emotional rating scale-second edition: Parent rating scale. *Children & Schools, 27*(3), 147-155.

Munger, R. L. (2000). Comprehensive needs-based assessment with adolescents. In W. E. Martin Jr. & J. L. Swartz-Kulstad (Eds.), *Person-environment psychology and mental health: Assessment and intervention* (pp. 11-38). Mahwah, NJ: Lawrence Erlbaum Associates.

Mushquash, C., & O'Connor, B. P. (2006). SPSS, SAS, and MATLAB programs for generalizability theory analyses. *Behavior Research Methods, 38*(3), 542-547.

Naumenko, O. (2015). Improving performance assessment score validation practices: An instructional module on generalizability theory. *Working Papers in Education, 1*(1), 1-17.

Nordness, P. D., Epstein, M. H., Cullinan, D., & Pierce, C. D. (2014). Emotional and behavioral screener: Test–retest reliability, inter-rater reliability, and convergent validity. *Remedial and Special Education, 35*(4), 211–217. https://doi.org/10.1177/0741932513497596

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Office of Special Education Programs. (2008). *Memorandum to chief state school officers*. Retrieved from https://sites.ed.gov/idea/files/policy_speced_guid_ idea_bapr_2008_bsppaprmemo081908.pdf 51 IDELR ¶ 49.

Office of Special Education Programs. (2011). *Memorandum to state directors of special education*. Retrieved from https://www2.ed.gov/policy/speced/ guid/idea/memosdcltrs/osep11-07rtimemo.pdf

Pallant, J. (2016). *SPSS survival manual: A step by step guide to data analysis using SPSS* (6th ed.). Maidenhead, England: Open University Press/McGraw-Hill.

Parker, R. I., Vannest, K. J., Davis, J. L., & Clemens, N. H. (2012). Defensible progress monitoring data for medium- and high-stakes decisions. *The Journal of Special Education, 46*(3), 141-151.

Parkes, J. (2000). Relationship between reliability and cost of performance assessment. *Education Policy Analysis Archives*, *8*(16), 1-14.

Pelham, W. E., Milich, R., Murphy, D. A., & Murphy, H. A. (1989). Normative data on the IOWA Conners Teacher Rating Scale, *Journal of Clinical Child Psychology, 18*(3), 259-262.

Peters, C. D., Kranzler, J. H., Algina, J., Smith, S. W., & Daunic, A. P. (2014). Understanding disproportionate representation in special education by examining group differences in behavior ratings. *Psychology in the Schools*, *51*(5), 452-465.

Pierce, C. D., Nordness, P. D., Epstein, M. H., & Cullinan, D. (2016). Applied examples of screening students at risk of emotional and behavioral disabilities. *Intervention in School and Clinic*, *52*(1), 6-11.

PRO-ED. *(*Firm). (2020). *About*. Retrieved from https://www.proedinc.com/

ProQuest. (Firm). (2016). *ProQuest statistical abstract of the United States 2016.* Lanham, MD: Bernan Press.

Pyrczak, F. (1973). Validity of the discrimination index as a measure of item validity. *Journal of Educational Measurement, 10,* 227-231.

Raines, T. C., Dever, B. V., Kamphaus, R. W., & Roach, A. T. (2012). Universal screening for behavioral and emotional risk: A promising method for reducing disproportionate placement in special education. *The Journal of Negro Education, 81*(3), 283-296.

Reynolds, C. R., & Kamphaus, R. W. (1992). *Manual for the behavior assessment system for children*. Circle Pines, MN: American Guidance Service.

Rowe, E. W., Curby, T. W., & Kim, H. (2019). Variance in teacher ratings of children's adjustment. *Journal of Psychoeducational Assessment,* 37*(1),* 26-39.

Ruble, L. A., McGrew, J. H., Wong, W. H., & Missall, K. N. (2018). Special education teachers' perceptions and intentions toward data collection. *Journal of Early Intervention, 40*(2), 177–191.

Salkind, N. J. (2010). *Encyclopedia of research design.* Thousand Oaks, CA: SAGE Publications, Inc.

Scardamalia, K., Bentley-Edwards, K. L., & Grasty, K. (2019). Consistently inconsistent: An examination of the variability in the identification of emotional disturbance. *Psychology in the Schools, 56*(4), 569-581.

Schumacker, R. E. (2010). *Standards for interpreting reliability coefficients*. Retrieved from http://appliedmeasurementassociates.com/ama/assets/File /Standards_for_Interpreting_Reliability_Coefficients.pdf.

Setyonugroho, W. (2017). Gentle introduction of generalizability theory analysis in OSCE using EduG for medical educators. *Journal of Computational and Theoretical Nanoscience, 23*(12), 12656-12659.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Los Angeles, CA: Sage.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922-932.

Sijtsma, K., & Van der Ark, L. A. (2015). Conceptions of reliability revisited, and practical recommendations. *Nursing Research, 64*(2), 128-136.

Smith, P. L. (1978). Sampling errors of variance components in small sample multifacet generalizability studies. *Journal of Educational Statistics, 3*(4), 319-346.

Smith-Millman, M. K., Flaspohler, P. D., Maras, M. A., Splett, J. W., Warmbold, K., Dinnen, H., & Luebbe, A. (2017). Differences between teacher reports on universal risk assessments. *Advances in School Mental Health Promotion, 10*(4), 235-249.

Snyder, T. D., & Dillow, S. A. (2015). *Digest of education statistics 2013*. Retrieved from https://nces.ed.gov/pubs2015/2015011.pdf

Splett, J. W., Smith-Millman, M., Raborn, A., Brann, K. L., Flaspohler, P. D., & Maras, M. A. (2018). Student, teacher, and classroom predictors of between-teacher variance of students' teacher-rated behavior. *School Psychology Quarterly, 33*(3), 460-468.

Tanner, N., Eklund, K., Kilgus, S. P., & Johnson, A. H. (2018). Generalizability of universal screening measures for behavioral and emotional risk. *School Psychology Review, 47*(1), 3-17.

Tedeschi, R. G., & Kilmer, R. P. (2005). Assessing strengths, resilience, and growth to guide clinical interventions. *Professional Psychology: Research and Practice, 36*(3), 230-237.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 8-14.

U.S. Department of Education. (2017). *IDEA part B child count and educational environments collection 2016-17*. Retrieved from https://www2.ed.gov/programs/osepidea/618-data/static-tables/index.html

U.S. Department of Education. (2018). *40th annual report to Congress on the implementation of the Individuals with Disabilities Education Act, 2018.* Retrieved from https://www2.ed.gov/about/reports/annual/osep/2018/parts-b-c/40th-arc-for-idea.pdf

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods, 23*(1), 1-26.

Volpe, R. J., & Briesch, A. M. (2016). Dependability of two scaling approaches to direct behavior rating multi-item scales assessing disruptive classroom behavior. *School Psychology Review, 45*(1), 39-52.

Volpe, R. J., & Briesch, A. M. (2017). Establishing evidence-based behavioral screening practices in U.S. schools. *School Psychology Review, 47*(4), 396-402.

Volpe, R. J., Briesch, A. M., & Gadow, K. D. (2011). The efficiency of behavior rating scales to assess inattentive–overactive and oppositional–defiant behaviors: Applying generalizability theory to streamline assessment. *Journal of School Psychology, 49*(1), 131-155.

Volpe, R. J., McConaughy, S. H., & Hintze, J. M. (2009). Generalizability of classroom behavior problem and on-task scores from the direct observation form. *School Psychology Review*, *38*(3), 382–401.

Volz-Sidiropoulou, E., Boecker, M., & Gauggel, S. (2016). The positive illusory bias in children and adolescents with ADHD: Further evidence. *Journal of Attention Disorders*, *20*(2), 178–186.

von der Embse, N. P., & Kilgus, S. P. (2018). Improving decision-making: Procedural recommendations for evidence-based assessment: Introduction to the special issue. *School Psychology Review, 47*(4), 329-332. doi:http://dx.doi.org.unco.idm.oclc.org/10.17105/SPR-2018-0059.V47-4

Webb, N. M., Rowley, G. L., & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development*, *21*, 81-90.

Webb, N. M., & Shavelson, R. J. (2005). Generalizability theory: Overview. *Encyclopedia of Statistics in Behavioral Science, 2,* 599-612.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). 4 reliability coefficients and generalizability theory. *Handbook of Statistics, 26*, 81-124.

Wehmeyer, M. L. (2019). *Strengths-based approaches to educating all learners with disabilities: Beyond special education*. New York, NY: Teachers College Press.

Wickard, G., & Hulac, D. (2017). Generalizability and dependability of a multi-item direct behavior rating scale in a kindergarten classroom setting. *Journal of Applied School Psychology*, *33*(2), 109-123.

Wixson, K. K., & Valencia, S. W. (2011). Assessment in RTI: What teachers and specialists need to know. *The Reading Teacher*, *64*(6), 466-469.

Wolcott, C. S., & Williford, A. P. (2018). Teacher and TA ratings of preschoolers' externalizing behavior: Agreement and associations with observed classroom behavior. *Topics in Early Childhood Special Education*, *34*(4), 211-222.

Wu, M., Tam, H. P., & Jen, T. (2016). *Educational measurement for applied researchers: Theory into practice*. Singapore: Springer.

Yell, M. L. (2012). *The law and special education* (3rd ed.). Upper Saddle River, NJ: Pearson.

APPENDIX A

PROPOSAL FOR PRO-ED REQUESTING ACCESS TO SECONDARY DATA

Proposal for PRO ED Requesting Access to Secondary Data

May 19, 2019

Dr. Epstein, Dr. Pierce, and Beth Allen, Director of Research for Pro-Ed,

Thank you for considering my request to access the norming data for the BERS -3. The data-based decision-making model used in schools relies on accurate and appropriate data as a crucial component in providing free appropriate public education for students with emotional disturbance. Validity and reliability coefficients for the subscales as well as the overall test for the BERS-2 are robust and indicate that the instrument measures what it purports to measure. Additionally, BERS is a measurement built using the federal operationalized definition of emotional disturbance. This is an important distinction as I seek to optimize decisions regarding special education and related services derived from collected data for this population of students.

The rationale with this study is to examine the magnitude of variance components attributable to different sources of error and to follow these findings with an investigation regarding how to make decisions in an applied setting. Classical testinG theory (CTT) forms the basis for determining reliability and validity coefficients for the BERS-3. However, a disadvantage of CTT is that it doesn't provide distinct sources of error, error from interactions of sources, and how these factors could impact the accuracy and precision of collected data used for data-based decision making. In an applied setting using different raters, different forms, and different occasions of measurement, for example, can create sources of error. Furthermore, measurements such as the BERS-3 seek to measure unobservable phenomenon, such as ability, rather than biological or physical traits. These two factors together indicate a need for the identification of the source of error so that measurement error may be regulated and accounted for.

Measurements such as the BERS-3 are useful in the provision of meaningful information about individuals by evaluating underlying traits. Rater, form, method, occasion, item, setting, and dimension are among the facets of generalizability that I would like to investigate. Using the norming data of the BERS-3 will allow me to answer the following questions:

1. How do generalizability or dependability coefficients change under different circumstances?
2. How do these changes inform how to best optimize the measurement?
3. How do the measurement administrative procedures need to be altered in order to achieve adequate levels of generalizability given constraints, such as number of raters, time, and cost?
4. Do the findings justify the application of generalizability theory and subsequent decision studies so that the appropriate special education and related services for students in the provision of free appropriate public education is met?

It would be of great value to me to have access to the norming data from the BERS-3 as I pursue my research interests through the dissertation phase of my doctoral program. Thank you again for your kind consideration.

Best,
Kara Scott Loftin

APPENDIX B

CORRESPONDENCE GRANTING PERMISSING FROM PRO-ED FOR SECONDARY DATA

Correspondence Granting Permission from PRO ED for Secondary Data

From: Pierce, Corey
Sent: Friday, May 10, 2019 12:13 PM
To: eallen@proedinc.com
Subject: Access to BERS-3 Norming Data

Hi Beth,
I have a doctoral student who is in the dissertation phase of her program. She has been actively assisting me with collection of data on both the BERS-3 and SAED-3 norming efforts. As you can read in the attached document that she prepared, Kara is interested in conducting some unique analyses on the BERS-3 data using G theory. I have spoken with Mike Epstein and he has given his verbal approval for Kara to access the data. He suggested I submit this request to you for approval from Pro-Ed. If you do approve, we would work with Jodie to get the raw data for Kara to analyze. Please let us know if you need any additional information for this request.
Have a nice weekend.
Sincerely,
Corey

Corey D. Pierce, Ph.D.
Associate Dean - College of Education and Behavioral Sciences
Professor - School of Special Education
University of Northern Colorado
McKee Hall #118
970-351-1662

This message originated from outside UNC. Please use caution when opening attachments or following links. Do not enter your UNC credentials when prompted by external links.

From: Pierce, Corey <Corey.Pierce@unco.edu>
Sent: Wednesday, May 22, 2019 10:54 AM
To: eallen@proedinc.com
Subject: RE: Access to BERS-3 Norming Data

Hi Beth,
I wanted to reach out and see if you were able to consider providing access to BERS-3 norming data for my doctoral student as described in the email below.
Thank you,
Corey

Corey D. Pierce, Ph.D.
Associate Dean - College of Education and Behavioral Sciences
Professor - School of Special Education
a

From: Elizabeth Allen <eallen@proedinc.com>
Sent: Wednesday, May 22, 2019 10:03:24 AM
To: Pierce, Corey
Cc: 'Jodie Martin'
Subject: RE: Access to BERS-3 Norming Data

Corey,

Thanks for the reminder. I am fine with your doc student working on this with your supervision. I have CC'd Jodie here so she knows it was ok'd with me.

Beth

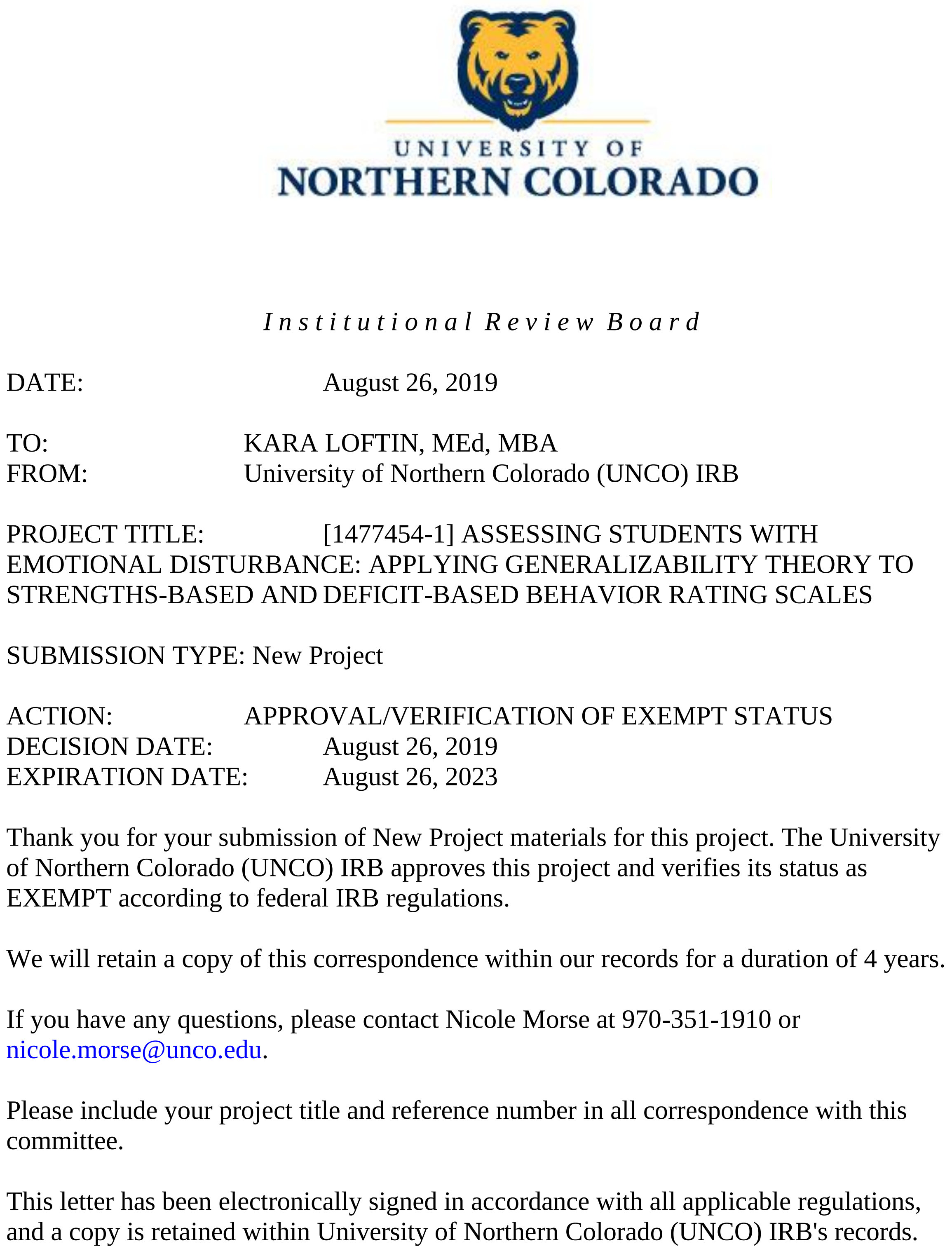
From: Pierce, Corey
Sent: Wednesday, May 22, 2019 10:25 AM
To: Loftin, Kara
Subject: Fwd: Access to BERS-3 Norming Data

Success!

Corey D. Pierce
Associate Dean
College of Education and Behavioral Sciences
Professor
School of Special Education
University of Northern Colorado
970-351-1662

APPENDIX C

INSTITUTIONAL REVIEW BOARD APPROVAL



UNIVERSITY OF
NORTHERN COLORADO

Institutional Review Board

DATE: August 26, 2019

TO: KARA LOFTIN, MEd, MBA
FROM: University of Northern Colorado (UNCO) IRB

PROJECT TITLE: [1477454-1] ASSESSING STUDENTS WITH EMOTIONAL DISTURBANCE: APPLYING GENERALIZABILITY THEORY TO STRENGTHS-BASED AND DEFICIT-BASED BEHAVIOR RATING SCALES

SUBMISSION TYPE: New Project

ACTION: APPROVAL/VERIFICATION OF EXEMPT STATUS
DECISION DATE: August 26, 2019
EXPIRATION DATE: August 26, 2023

Thank you for your submission of New Project materials for this project. The University of Northern Colorado (UNCO) IRB approves this project and verifies its status as EXEMPT according to federal IRB regulations.

We will retain a copy of this correspondence within our records for a duration of 4 years.

If you have any questions, please contact Nicole Morse at 970-351-1910 or nicole.morse@unco.edu.

Please include your project title and reference number in all correspondence with this committee.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within University of Northern Colorado (UNCO) IRB's records.