The Effects of Rater Training on Rater Effects and Validity of Direct Behavior Ratings

Abigail E. Pruitt
abigailxpruitt@gmail.com

Follow this and additional works at: https://huskiecommons.lib.niu.edu/allgraduate-thesesdissertations

Recommended Citation
https://huskiecommons.lib.niu.edu/allgraduate-thesesdissertations/7569

This Dissertation/Thesis is brought to you for free and open access by the Graduate Research & Artistry at Huskie Commons. It has been accepted for inclusion in Graduate Research Theses & Dissertations by an authorized administrator of Huskie Commons. For more information, please contact jschumacher@niu.edu.
ABSTRACT

THE EFFECTS OF RATER TRAINING ON RATER EFFECTS AND VALIDITY OF DIRECT BEHAVIOR RATINGS

Abigail Pruitt, M.A.
Department of Psychology
Northern Illinois University, 2022
Kara Styck, Director

Educators have a responsibility to accurately measure student behavior in order to identify students in need of additional behavioral support. Current behavior screening tools can be lengthy or difficult to complete, and direct behavior ratings (DBRs) offer a solution. However, DBRs are rater-mediated assessments, prone to rater effects. Rater training methods can be used to mitigate these rater effects, but previous research has not investigated the best training method for reduction of rater effects. Additionally, many facet Rasch measurement (MFRM) provides an opportunity to adjust student ratings in response to individual rater tendencies of severity and leniency. Therefore, the primary purpose of the study was to investigate the optimal training method for the reduction of rater effects in DBRs. A secondary purpose was to identify the training method that produced the highest level of external criterion validity between DBRs and systematic direct observations (SDOs). Results indicated that rater training reduced rater effects for DBRs on Disruptive Behavior but did not significantly reduce rater effects for DBRs on academic engagement. Additionally, rater training produced the highest levels of validity across all DBRs. Implications for practice include providing training to raters and using MFRM to identify problematic DBR items and achieve high DBR validity.
THE EFFECTS OF RATER TRAINING ON RATER EFFECTS AND VALIDITY OF DIRECT BEHAVIOR RATINGS

BY

ABIGAIL PRUITT
©2022 Abigail Pruitt

A THESIS SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE
MASTER OF ARTS

DEPARTMENT OF PSYCHOLOGY

Thesis Director:
Kara M. Styck
# TABLE OF CONTENTS

| LIST OF TABLES | iv |
| LIST OF APPENDICES | v |

Chapter

<p>| I. INTRODUCTION | 1 |
| II. REVIEW OF THE LITERATURE | 5 |
| Direct Behavior Ratings | 5 |
| Definition | 5 |
| DBR-SIS | 6 |
| DBR-MIS | 8 |
| Rater Effects | 9 |
| Rater Effects on Direct Behavior Ratings | 10 |
| ANOVA | 11 |
| Generalizability Theory | 13 |
| Many-Facet Rasch Measurement | 16 |
| Rater Training | 17 |
| The Current Study | 18 |
| Research Questions and Hypotheses | 19 |
| III. METHODOLOGY | 21 |
| Participants | 21 |
| Measures | 23 |
| DBR-MIS | 23 |
| SDO | 23 |
| Procedure | 24 |
| Rating Design and Procedure | 24 |
| Independent Variable | 25 |
| Dependent Variables | 27 |
| Data Analysis Plan | 29 |
| IV. RESULTS | 31 |
| Question 1 | 31 |
| Academic Engagement | 31 |
| Disruptive Behavior | 33 |
| Question 2 | 36 |</p>
<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Pearson Correlations Between DBRs and SDOs Disaggregated by Training Status</td>
<td>36</td>
</tr>
</tbody>
</table>
LIST OF APPENDICES

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. DIRECT BEHAVIOR RATINGS</td>
<td>49</td>
</tr>
<tr>
<td>B. TRAINING MODULE</td>
<td>53</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

Disruptive behaviors (i.e., tantrums, oppositional defiance, eloping, physical or verbal violence) and academic enablers (i.e., engagement, motivation, interpersonal skills, and study skills) are commonly targeted for social-emotional and behavioral assessment and intervention within multi-tiered systems of support (Chafouleas, Christ & Riley-Tillman, 2009; Manasiev et al., 2019; Volpe & Briesch, 2012). Students who display disruptive behaviors in the classroom tend to perform lower than their peers on tests of academic achievement (Hinshaw, 1992; Nelson et al., 2004) and are at higher risk for suspension in middle school and dropout in high school (Jimerson et al., 2000; Raffaele Mendez, 2003). This decreases their likelihood of college admission and limits their lifelong employment opportunities (U.S. Bureau of Labor Statistics, 2020). In contrast, students who display academic enablers tend to excel in reading and mathematics (DiPerna et al., 2002, 2005) and demonstrate lower rates of disruptive behaviors (Kretschmer et al., 2014; Thompson et al., 2011). As a result, the accurate identification of children who need additional support targeting these social-emotional and behavioral skills is a critical component of effective intervention within a multi-tiered system of supports (MTSS).

According to the results of Benson et al. (2019), social-emotional and behavioral skills are typically screened within MTSS using one of three assessment methods: (a) systematic direct
observation (SDO), (b) teacher behavior rating scales (BRSs), and (c) direct behavior ratings (DBRs). SDOs require a teacher, paraprofessional, or other educational support staff to observe student behavior in real time and collect descriptive data on a specified behavioral dimension (frequency, duration, latency). SDOs are considered the “gold standard” for behavior assessment due to the highly accurate information they provide (Briesch et al., 2010; Chafouleas, Jaffery, et al., 2013; Chafouleas et al., 2005; C. Riley-Tillman et al., 2008; T. C. Riley-Tillman et al., 2009). This assessment method is highly precise when compared to other assessments because student behavior is measured directly and SDOs can provide more specific information about specific students’ strengths and areas of improvement, but SDOs are difficult to complete in a classroom setting because they require additional personnel to conduct the direct observation while a teacher teaches.

Teacher BRSs require a teacher to rate students’ behavior over a specified period of time (e.g., in the last several months) using a Likert-type scale. For example, the teacher report form of the Social, Academic, and Emotional Behavior Risk Screener (SAEBRS; Kilgus et al., 2013) asks teachers to rate student behavior on 19 items “over the past month” on a 4-point scale ranging from Never to Almost Always (Kilgus et al., 2013). Student ratings are then aggregated across items to derive three subscale scores designed with the intention of measuring students’ social, academic, and emotional functioning. Teacher BRSs are easy to complete; however, even screeners can be lengthy. For example, the SAEBRS takes one to three minutes to complete according to the test publisher, which would require 20 minutes to 1 hour of teacher time for a class of 18 students (Kilgus et al., 2013). Teacher BRSs also require teachers to rely on memory
to rate the frequency/intensity of student behavior tendencies over the specified time period, which introduces an additional source of measurement error related to the teacher rater.

Finally, DBRs require a teacher to rate a specific target behavior on a Likert-type scale after directly observing a student for a specified period of time (e.g., 10 minutes). DBRs are hypothesized to combine the strengths of teacher BRSs (i.e., feasibility of completion) and SDOs (i.e., reliable/valid behavior descriptions; Volpe & Briesch, 2012); they are highly flexible in that they can be adapted to measure any social-emotional target behavior. Moreover, there is some evidence that DBRs have high teacher acceptability and that teachers may prefer DBRs over other social-emotional and behavioral assessment methods (Chafouleas, Kilgus, & Hernandez, 2009). Further, DBRs can be completed up to 6 hours after a given observation with no significant effects on accuracy (Dart et al., 2017). However, they are an indirect assessment method and share the limitations of teacher BRSs as a result.

Specifically, DBRs are *rater-mediated* assessments. A rater-mediated assessment involves a rater who judges an examinee’s proficiency on a certain construct (e.g., academic engagement) based on the rater’s understanding of that construct (Eckes, 2009). This means a rater is providing a score based on both their perception of an examinee’s proficiency level as well as their own understanding of a given construct. Therefore, rater-mediated assessments are prone to *rater effects*. Rater effects can be described as variability in ratings that is attributable to the individual rater rather than the examinee. Rater effects are often defined in terms of the severity (tendency to rate low) or leniency (tendency to rate high) of the rater, but they can also include central tendency, restriction of range, and halo/horns effects (Myford & Wolfe, 2003, 2004).
Unfortunately, rater effects have been found to account for a substantial amount of variance in DBR scores (Briesch et al., 2010; Chafouleas, Kilgus, et al., 2013; Christ et al., 2010; Hartman et al., 2007; Peters et al., 2014; T. C. Riley-Tillman et al., 2009; Smith-Millman et al., 2017; Splett et al., 2018; Volpe & Briesch, 2012, 2016). This variance attributable to raters ranges from 2% in cases of raters trained on DBRs to 20% for untrained raters (Briesch et al., 2010; Wickerd & Hulac, 2017), which suggests that rater training is an important consideration in addressing rater effects on DBRs. Furthermore, there is some emerging evidence that rater effects on DBRs may be especially pronounced for students’ who need the most support in the classroom (i.e., students who display more frequent problem behaviors and less frequent academic engagement/motivation; Anthony et al., accepted). This could result in an inefficient allocation of school resources and hinder student success. In particular, if a student were rated by an especially lenient rater, they may not be flagged as needing additional support. Conversely, if a student were rated by an especially severe rater, they may be flagged as needing additional support when their social-emotional and behavioral functioning is not actually at risk. As a result, the overall aims of the present study are to 1) determine the degree to which rater training reduces rater effects on DBR scores and 2) examine the impact of rater effects on the validity of DBR scores.
CHAPTER II

REVIEW OF THE LITERATURE

Direct Behavior Ratings

Definition

DBRs have the potential to be used widely in schools for social-emotional and behavioral assessment given that they offer a balance between precision and feasibility (e.g., length, ease of use) when compared to other social-emotional and behavioral assessment methods (Christ et al., 2009). Historically, DBRs have been primarily used in schools to screen students for social-emotional and behavioral problems in order to identify individual students in need of additional social-emotional and behavioral support (Chafouleas, Christ, & Riley-Tillman, 2009; Chafouleas, Kilgus, et al., 2009, 2013; Wickerd & Hulac, 2017). However, there is some preliminary evidence from single-case design research supporting the use of DBRs for formative assessment in progress monitoring applications, as well (Chafouleas et al., 2012; T. C. Riley-Tillman et al., 2009). DBRs come in one of two forms: single-item scales (DBR-SISs) and multiple-item scales (DBR-MISs). A DBR-SIS form contains a single item and asks a teacher to rate a student’s proficiency on a given domain of social-emotional and behavioral functioning.
using a single item. For example, a DBR-SIS form may ask a teacher to “circle the letter [rating] that best describes the student’s behavior during the observation period” for a given domain on a 6-point scale ranging from *Never* to *Always* (Volpe & Briesch, 2012). In contrast, a DBR-MIS form contains multiple items that assess a given domain of social-emotional and behavioral functioning. For example, a teacher may be asked to rate a student on five items that assess specific components of a student’s level of academic engagement on a 6-point scale ranging from *Never* to *Always* (e.g., “Finishes work on time,” “Actively participates in class”; Volpe & Briesch, 2012).

**DBR-SIS**

DBR-SISs preceded DBR-MISs in development and have garnered a wide body of psychometric literature. DBR-SIS ratings have generally demonstrated moderate to strong relations with scores from teacher BRSs (Chafouleas, Kilgus, & Hernandez, 2009; Kilgus et al., 2012) and SDOs (Smith et al., 2018) in elementary student samples. For example, results of empirical studies have reported that DBR-SIS ratings of academic engagement have significant positive associations with the Social Skills Rating System (Gresham & Elliott, 1990) Academic Competence ($r = .53$) and Social Skills ($r = .86$) scales and that DBR-SIS ratings of academic engagement have significant, positive associations with the Social Skills Improvement System (Elliott & Gresham, 2007) Motivation to Learn ($r = .77$) and Prosocial Behavior scales ($r = .67$; Kilgus et al., 2012). In addition, results of empirical studies have reported that DBR-SIS ratings of academic engagement have significant, negative associations with the Social Skills Rating
System Problem Behaviors scale ($r = -.88$; Chafouleas, Kilgus, & Hernandez, 2009) and strong negative associations with the Behavioral and Emotional Screening System (Kamphaus & Reynolds, 2007) overall composite ($r = -.77$; Kilgus et al., 2012). Finally, DBR-SIS ratings for academic engagement have also demonstrated moderate to strong relations with SDO scores (Smith et al., 2018). Specifically, DBR-SIS ratings were demonstrated to have significant, positive associations with SDO ratings during the baseline phase ($r = .44, p < .05$) and intervention phase ($r = .60, p < .01$) of a SDO protocol (Smith et al., 2018).

In addition to academic engagement, DBR-SIS ratings of disruptive behavior have also demonstrated moderate to strong relations with scores from teacher BRSs and SDOs. For example, DBR-SIS ratings for disruptive behavior have demonstrated significant, positive associations with the Problem Behaviors scale ($r = .82$; Chafouleas, Kilgus, & Hernandez, 2009) as well as strong positive associations ($r = .66$) with the Behavioral and Emotional Screening System (Kamphaus & Reynolds, 2007; Kilgus et al., 2012) in elementary-student samples. Additionally, DBR-SIS ratings for disruptive behavior have also demonstrated significant, positive correlations with SDO ratings during a baseline data collection phase ($r = .87, p < .01$) as well as nonsignificant, positive correlations during an intervention phase ($r = .29$; Smith et al., 2018) for an elementary sample. Further, DBR-SIS ratings for disruptive behavior have demonstrated significant, negative correlations with the Academic Competence ($r = -.42$) and Social Skills ($r = -.77$) scales on the Social Skills Rating System (Chafouleas, Kilgus, & Hernandez, 2009), as well as the Social Skills Improvement System—Motivation to Learn index ($r = -.606$) and the Social Skills Improvement System—Prosocial Behavior index ($r = -.541$; Kilgus et al., 2012) for elementary-student samples.
Finally, DBR-SIS ratings for compliance have demonstrated significant correlations with BRS. For example, DBR-SIS ratings for compliance demonstrated strong positive associations ($r = -.66$) with the Behavioral and Emotional Screening System (Kamphaus & Reynolds, 2007; Kilgus et al., 2012). DBR-SIS ratings for compliance have demonstrated significant, positive correlations with the Social Skills Improvement System--Prosocial Behavior index ($r = .520$) and the Social Skills Improvement System--Motivation to Learn index ($r = .659$; Elliott & Gresham, 2007; Kilgus et al., 2012).

**DBR-MIS**

Due to its more recent development, there have been significantly fewer investigations considering the psychometric properties of DBR-MISs. However, DBR-MISs for disruptive behavior ratings have demonstrated small to moderate, non significant, positive associations ($r = .19$ to .37) with scores from the Attention-Deficit Hyperactivity Disorder Symptom Checklist – Fourth Edition (Gadow & Sprafkin, 1997) with a sample of students in kindergarten ($n = 18$; Wickerd & Hulac, 2017).

The feasibility of DBRs combined with these psychometric properties have led to the recommendation of DBR use among elementary and middle grades with various cut score recommendations for the provision of services (Chafouleas, Kilgus, et al., 2013). However, results of generalizability studies have demonstrated that it can require as many as 11 assessment occasions to reach an acceptable level of reliability of .80 for educational programming with DBR-MISs (Volpe & Briesch, 2012) and over 100 assessment occasions to reach this same level
of acceptable reliability with DBR-SISs (Volpe & Briesch, 2012). This need for many rating occasions is problematic for educators since screening decisions are typically made on the basis of a single assessment score, and an accurate and efficient assessment method is needed to provide students with services (Christ et al., 2009). However, research has shown that the primary source of error in DBRs is the raters themselves (Briesch et al., 2010; Chafouleas, Christ, & Riley-Tillman, 2009; Chafouleas, Kilgus, et al., 2013; Christ et al., 2010; Volpe & Briesch, 2012, 2016), and a few statistical methods have been proposed to address this issue.

Rater Effects

The term “rater effects” describes variability in ratings that is due to characteristics of the rater rather than the examinee. Rater effects include the severity/leniency of a rater, halo/horns effects, restriction of range, and central tendency (a special case of restriction of range) and can characterize the rating behavior of a group of raters or individual raters’ rating tendencies (Myford & Wolfe, 2003, 2004). Severity/leniency describes the tendency of a rater to rate an examinee as having lower/higher proficiency on a target behavior (e.g., academic engagement/motivation) when compared to other raters (Eckes, 2009). For example, a severe rater may rate a student as having very low academic engagement while other raters rate that same student as having a moderate or high level of academic engagement, as would be the case for a lenient rater. The halo/horns effect describes the tendency of a rater to provide similar ratings for an examinee across all items based on some perceived characteristic of the examinee (Myford & Wolfe, 2003, 2004). Consequently, halo/horns effects result in a student being rated...
universally low/high regardless of the rated domain. In contrast, restriction of range describes a rater’s tendency to overuse a select portion of a rating scale, rather than the entire scale range (Myford & Wolfe, 2003, 2004). For example, a rater may only assign ratings of Often, Very Often, or Always to each student they are assigned to rate, thereby neglecting ratings of Never, Rarely, and Sometimes. Finally, the central tendency effect is a special case of restriction of range that describes a rater’s tendency to rate examinees using only the midpoint and surrounding rating categories rather than the entire scale range, thus rating most (or all) students as having an “average” level of proficiency on the target behavior (Myford & Wolfe, 2003, 2004).

Rater effects are problematic because they can cause spuriously high or attenuated ratings that reflect characteristics about the rater in addition to the actual performance of the student being rated. This can invalidate social-emotional and behavioral assessment data and ultimately lead to withholding support from students who do need it and providing additional support to students who do not.

Rater Effects on Direct Behavior Ratings

Researchers have sought to investigate and address rater effects on DBRs using two types of methods: statistical techniques and rater training. A number of statistical techniques have been used to investigate and characterize rater effects on DBRs, including analysis of variance (ANOVA; Chafouleas, Jaffery, et al., 2013; Dart et al., 2017), generalizability theory (GT; Briesch et al., 2010; Chafouleas & Riley-Tillman, 2010; Chafouleas, Christ, et al., 2009; Christ et al., 2010; Daniels et al., 2017; Leposa, 2017; Volpe & Briesch, 2012, 2016; Wickerd & Hulac,
and many-facet Rasch measurement (MFRM; Anthony et al., accepted). Rater training techniques have varied in previous studies, including no training (Chafouleas, Kilgus, et al., 2013, p. 201; Daniels et al., 2017; Volpe & Briesch, 2012); providing a detailed definition with example behaviors (Briesch et al., 2010; Chafouleas, Christ, & Riley-Tillman, 2009; Christ et al., 2010; Dart et al., 2017); and providing a definition, example ratings, and opportunities for practice with feedback (Chafouleas et al., 2010; Leposa, 2017; Volpe & Briesch, 2016; Wickerd & Hulac, 2017). Each of these methods represents a different means of investigating rater effects on DBRs and each method provides unique information.

**ANOVA**

Dart et al. (2017) used ANOVA to investigate the degree to which the length of time between a direct observation and the completion of a DBR impacts rating accuracy (i.e., the discrepancy between DBR ratings and SDO ratings). A sample of 241 undergraduate students were first shown a training video on how to complete DBRs and were then provided with a definition for the behavior of interest: out-of-seat behavior. Participants were then instructed to watch a 10-minute video segment of a third-grade classroom, focusing on the behavior of a single White, male student. After viewing the video, the undergraduates were divided into eight latency groups. The first group completed a DBR-SIS immediately after viewing the video, and the other seven groups waited 5 minutes, 15 minutes, 30 minutes, 1 hour, 2 hours, 4 hours, and 6 hours, respectively, before completing the DBR-SIS. Results indicated that there was no significant relation between completion latency and DBR-SIS rating accuracy (Dart et al., 2017).
However, while these analyses indicated that the two groups demonstrated similar means and levels of variance, they did not identify how much of the variance in DBR-SIS ratings was attributable to the individual raters whether as a function of observation rating latency.

In a similar study utilizing ANOVA, 113 undergraduate students rated videos of elementary-aged students on either academic engagement or disruptive behavior (Chafouleas, Jaffery, et al., 2013). Here, researchers were interested in the effect of positively or negatively worded target behaviors (e.g., “respectful” vs. “disrespectful”) on rater bias for DBR-SISs compared to both SDO ratings and expert DBR-SIS raters. Expert ratings were obtained via a consensus between 13 research personnel with extensive DBR training, including professors, postdoctoral fellows, and graduate students (Chafouleas, Jaffery, et al., 2013). *Rater error* was defined as the absolute difference between mean participant DBR ratings and the criterion measures (i.e., SDO ratings and expert DBR ratings). Further, *rater bias* was defined as the difference between mean participant DBR ratings and the criterion measures (i.e., SDO ratings and expert DBR ratings). The distinction between rater error and rater bias in this study was that rater bias provided information about whether the raters, as a group, tended to over- or underestimate the behaviors they were rating, whereas rater error did not (Chafouleas, Jaffery, et al., 2013). The undergraduate students were provided with definitions for the target behaviors and were then shown a series of nine 60-second clips of elementary students in the classroom before filling out a DBR-SIS for each student. As calculated by using SDO and expert ratings, researchers found that rater bias was present to some degree for all target behaviors, with greater bias present for constructs involving respectful behavior (Chafouleas, Jaffery, et al., 2013).
This method of analysis allowed researchers to compare accuracy of DBR-SIS ratings compared to expert and SDO ratings, but it did not provide information regarding individual raters’ rating tendencies. Additionally, the way in which the researchers measured both rater error and rater bias poses limitations on the interpretation of the results. Specifically, the researchers compared participant DBR ratings to “expert” ratings, treating the expert ratings as objectively correct. However, because every rater demonstrates some degree of severity/leniency in their ratings when compared to others, an absolute “true” rating for a behavior is not possible to obtain.

**Generalizability Theory**

Generalizability theory (GT) is another statistical method, derived from ANOVA, that allows researchers to determine the amount of variance attributable to various sources of error (Leposa, 2017). Researchers have employed GT to improve upon ANOVA results by estimating how much variance in a set of ratings is due to the rater, rather than the examinee, on both DBR-SIS and DBR-MIS (Briesch et al., 2010; Chafouleas et al., 2010; Chafouleas, Christ, & Riley-Tillman, 2009; Christ et al., 2010; Volpe & Briesch, 2012, 2016; Wickerd & Hulac, 2017).

Researchers have employed GT to investigate the amount of variance attributable to raters when using DBR-SIS in multiple studies (Briesch et al., 2010; Chafouleas et al., 2010; Chafouleas, Christ, & Riley-Tillman, 2009; Volpe & Briesch, 2012). For example, 10% of the variance in DBR scores was attributable to raters when graduate students were not provided training before using a DBR for disruptive behavior (Volpe & Briesch, 2012). Other studies have investigated the amount of variance attributable to raters when the raters are provided minimal
training, such as the definition of the target behavior (Briesch et al., 2010; Chafouleas, Christ, & Riley-Tillman, 2009; Christ et al., 2010). These studies include raters that were undergraduates rating videos of students (Chafouleas, Christ, & Riley-Tillman, 2009; Christ et al., 2010) and teachers rating students in-person (Briesch et al., 2010). In each case, raters accounted for 17-20% of the variance in DBR scores (Briesch et al., 2010; Chafouleas, Christ, & Riley-Tillman, 2009; Christ et al., 2010). Finally, there have been a few studies that examined the impact of raters on DBRs when the raters had been given higher levels of training, such as a target behavior definition with feedback on practice ratings (Chafouleas et al., 2010; Leposa, 2017). In one of these studies, two teachers and two paraprofessionals rated nine kindergarten students in person on academic engagement and disruptive behavior, and the Person by Rater by Day interaction accounted for 17% of the variance in scores (Leposa, 2017). In contrast, the second study included two research assistants and two teachers who rated seven eighth-grade students in person on these same target behaviors and a relatively small amount of variance was found to be attributable to the rater (5% for academic engagement and 2% for disruptive behavior, respectively; Chafouleas et al., 2010). Taken together, the current research provides mixed findings regarding the effectiveness of training on minimizing rater effects.

Most literature has investigated rater effects on DBR-SISs. However, some studies have investigated rater effects on DBR-MISs as well. Rater effects appear to be somewhat smaller on DBR-MISs when compared to DBR-SIS (Volpe & Briesch, 2012), which is largely due to the increase in reliability afforded by the increase in the number of items rated. For example, in one study two graduate student observers were provided with no training before being tasked with watching a video of eight seventh-grade students and rating the students’ behavior on a
DBR-MIS for academic engagement and disruptive behavior. Results indicated that 4% of the variance in DBR-MIS scores was attributable to the raters for each target behavior (Volpe & Briesch, 2012). In another study, seven graduate student observers were trained to use a DBR-MIS for disruptive behavior by being provided with both a definition of the target behavior and feedback on practice ratings. Observers rated nine seventh-grade students, and results indicated that variance in DBR-MIS scores attributable to the raters was similarly small, 4% (Volpe & Briesch, 2016). Finally, in a third study, a special education teacher and a paraprofessional were trained to use DBR-MISs for disruptive behavior and observed students in a live kindergarten classroom setting ($n = 18$; Wickerd & Hulac, 2017). These two educators entered the classroom at the same time every day to observe the students and perform the ratings. The percent of variance in DBR-MIS scores attributable to raters was negligible ($< 2$%; Wickerd & Hulac, 2017).

Although at first glance it appears that rater effects may not be a serious issue for DBR-MIS, GT only provides a means by which rater effects can be isolated, identified, and differentiated from other sources of error on DBR-MISs. However, GT does not offer a method for controlling for or adjusting DBR-MIS scores to account for rater effects on individual students’ DBR-MIS scores. In addition, GT does not provide nuanced information about observers’ individual rating tendencies (e.g., severity/leniency, restriction of range, central tendency) that may indicate a need for additional training for specific observers or identify problems with individual DBR-MIS items that may be contributing to rater effects.
Many-facet Rasch measurement (MFRM; Linacre, 1989) is a statistical technique within the scaling tradition that can be used to isolate, identify, differentiate, and control for variance attributable to raters on a rater-mediated assessment, such as a DBR-MIS. Anthony and colleagues (accepted) used both GT and MFRM in order to evaluate the presence of rater effects on DBR-MIS for a sample of 126 elementary-school students. Students were rated by four raters on two separate DBR-MISs targeting academic engagement/motivation and disruptive behaviors. The raters in the study included four graduate students, each of whom completed 10-minute in vivo observations and ratings for 22 different students contemporaneously and an additional 108 ratings on a different set of students independently. Results indicated that raters accounted for 6.6% of the variance in DBR-MIS scores on Academic Engagement/Motivation and 5.7% of the variance in DBR-MIS scores on Disruptive Behavior, similar to prior findings published using GT methodology. However, results of MFRM analyses provided a more nuanced picture of rater effects on the two DBR-MIS. For example, results indicated that raters significantly differed from one another with regard to tendencies towards severity/leniency and that these differences adversely impacted individual students’ scores. In addition, results indicated that raters applied rating categories in an incohesive manner on the Academic Engagement/Motivation DBR-MIS, rendering the ratings unusable (higher ratings did not correspond to more student academic engagement/motivation). Finally, the authors found that raters demonstrated “restriction of range” in their ratings, as a group, on the Disruptive Behaviors DBR-MIS whereby ratings of
disruptive behavior were largely concentrated at the bottom end of the rating scale (i.e., ratings of *Never* and *Rarely*).

**Rater Training**

Finally, rater training has also been used to reduce rater effects with the goal being to prevent rater effects altogether. The literature can be categorized into three different approaches for rater training: providing DBR definitions as well as feedback to practice items, providing DBR definitions without rater feedback, and no training at all. A few studies have provided intensive DBR training prior to rating occasions (i.e., DBR definitions with rater feedback), and these studies have shown decreased levels of rater effects as compared to studies in which the raters received little or no prior training on DBRs (Briesch et al., 2010; Chafouleas et al., 2010; Chafouleas, Christ, & Riley-Tillman, et al., 2009; Christ et al., 2010; Volpe & Briesch, 2012; Wickerd & Hulac, 2017). However, similar studies have demonstrated conflicting results, such that high levels of rater training do not adequately mitigate rater effects, with a significant amount of variance still attributable to rater interactions (Leposa, 2017; Volpe & Briesch, 2016). In a study conducted by Taipalus and colleagues (2021), raters who completed an online training module prior to completing DBRs demonstrated generally higher accuracy than those who did not complete the online training. However, a visual analysis of these results suggested that the impact of the training was virtually negligible, suggesting that training may not be an efficient means for reducing rater effects on DBRs.
The current study extends the research on DBR-MIS by investigating the impact of rater training on rater effects. Previous investigations have demonstrated that rater effects exist in large magnitudes for DBRs on elementary samples (Briesch et al., 2010; Chafouleas, Christ, & Riley-Tillman, 2009; Volpe & Briesch, 2016). Between 4 - 20% of variance in student scores has been found attributable to the rater completing the DBR rather than actual differences in student behavior (Briesch et al., 2010; Chafouleas, Christ, & Riley-Tillman, 2009; Chafouleas, Jaffery, et al., 2013; Christ et al., 2010; Volpe & Briesch, 2012). These effects were found to be present in studies with rater training (Briesch et al., 2010; Chafouleas, Christ, & Riley-Tillman, 2009; Christ et al., 2010) and in studies without rater training (Chafouleas, Jaffery, et al., 2013; Volpe & Briesch, 2012). No study, however, was successful in completely eliminating rater effects through rater training. While various methods of rater training have been conducted in the literature, a comparative approach has not yet been taken when considering the impact of rater training on rater effects. Thus, the primary purpose of this study was to determine the impact of rater training compared to no rater training on rater effects.

A secondary purpose of this study was to determine the degree to which rater effects impact DBR validity. Previous studies have utilized SDOs as a “gold standard” measurement tool for evaluating student behavior, and relations between DBRs and SDOs for concurrent validity have also been previously examined (Briesch et al., 2010; Chafouleas, Jaffery, et al., 2013; Chafouleas et al., 2005; Riley-Tillman et al., 2008; Riley-Tillman et al., 2009). Therefore, SDOs were used as the external criterion validity measure to support the use of DBRs as
Rater-mediated assessments. Rater effects, by definition, are sources of variance that are attributable to the rater facet rather than the examinee facet (Eckes, 2009; Myford & Wolfe, 2003, 2004). The current study also sought to investigate the degree to which elevated levels of rater effects impact DBR and SDO concurrent validity.

Research Questions and Hypotheses

Question 1: Do rater effects reduce the validity of DBR scores?

Prediction 1: Yes. Because rater effects indicate a percentage of variance that is attributable to individual differences in raters rather than differences in examinee performance (Eckes, 2009; Myford & Wolfe, 2003, 2004), rater effects reduce the validity of DBR scores. This can be seen through measures of concurrent validity, such as the relation between DBR scores and SDOs (Briesch et al., 2010; Chafouleas, Jaffery, et al., 2013; Chafouleas et al., 2005; Riley-Tillman et al., 2008; Riley-Tillman et al., 2009). Scores with lower levels of rater effects are anticipated to have stronger relations with SDO data, demonstrating greater concurrent validity.

Question 2: Does rater training significantly reduce rater effects when compared to no rater training?

Prediction 2: Yes, rater training will significantly reduce differences in rater severity/leniency on DBR-MISs as compared to the effects of raters who did not receive training. In prior studies, raters tend to show higher levels of rater effects when given no training (Volpe & Briesch, 2012) or minimal training (Briesch et al., 2010; Chafouleas, Christ, Riley-Tillman, 2009; Christ et al.,
Raters who receive training on DBRs prior to the rating occasion tend to exhibit fewer rater effects (Chafouleas et al., 2010; Leposa, 2017). It is therefore expected that raters in the training group of the current study will demonstrate lower levels of rater effects as compared to raters in the no training group.
CHAPTER III

METHODOLOGY

Participants

Graduate students \( (N = 10) \) at a university in the midwestern United States observed and rated the disruptive behavior and academic engagement of a sample of 15 middle school students from video recordings of a general education classroom engaged in a math lesson. This graduate student sample size is similar to those of previous studies specifically investigating rater effects, which ranged from sample sizes of two to six (Briesch et al., 2010; Chafouleas et al., 2010; Taipalus et al., 2021; Volpe & Briesch, 2012, 2016; Wickerd & Hulac, 2017). Graduate student observers eligible to participate in the study included students enrolled in the school psychology program of study at the author’s institution who had had little or no prior training on DBRs. Observers \( (N = 10) \) included three male and seven female graduate students, of which 30% were in their first year of graduate study, 10% were second years, 20% were third years, 20% were fourth years, and 10% were in their fifth year or beyond of graduate study. The highest degree completed for observers included a bachelor’s degree (50%), a master’s degree (40%), and a specialist or professional diploma (10%). Racial demographics for observers included 80% White, 10% Asian, and 10% Other; 20% of observers identified as Hispanic or Latino. When
surveyed about previous experiences with DBRs, 70% of observers reported that they had
learned about DBRs in previous coursework and 30% of observers reported that they had no
prior experience with DBRs. Of the 70% who reported that they had learned of DBRs in
previous coursework, two observers had also read about DBRs through independent study and
one observer reported that they have used DBRs in applied experiences. Graduate student
observers were incentivized to participate in the study by being permitted to count their time
spent rating videos towards their program-required hours of observing students in a school
setting.

Videos of elementary students were derived from Volpe and Briesch (2012). The video
clips used in the present study included eight male and nine female seventh-grade students in a
general education math class at an urban public charter school in the northeastern region of the
U.S. All of the students enrolled at the school were students of color, with over 70% of the
student body eligible for free and reduced-price lunch (Volpe & Briesch, 2012). Each video was
approximately 12 minutes in length, which is similar to observation periods in prior studies
(Chafouleas et al., 2010; Dart et al., 2017; Volpe & Briesch, 2012, 2016). For further details on
participant demographics, see Volpe and Briesch (2012).
Measures

**DBR-MIS**

DBR-MISs targeting academic engagement/motivation and disruptive behavior were used in the current study following the procedures outlined in Volpe and Briesch (2012). Each DBR-MIS contained instructions for observers to rate the frequency of the target behavior during the observation window on a 7-point ordinal scale ranging from 1 (i.e., the behavior *Never* occurred) to 6 (i.e., the behavior *Always* occurred). Disruptive behavior includes 1) “Clowns around,” 2) “Noisy,” 3) “Out of seat/area,” 4) “Talks to classmates when inappropriate,” and 5) “Calls out.” Academic engagement/motivation behaviors include 1) “Finishes work on time,” 2) “Actively participates in class,” 3) “Stays on task,” 4) “Works independently,” and 5) “Starts tasks promptly.” Volpe and Briesch (2012) reported reliability-like coefficients for each of the DBR-MISs used in this study. These coefficients fell in the acceptable range for both Academic Engagement/Motivation ($r^2 = .85$, $F = .82$) and Disruptive Behavior ($r^2 = .64$, $F = .63$).

**SDO**

The current study included a partial interval recording SDO at 15-second intervals completed by two trained undergraduate students. Partial interval recording is more appropriate for capturing low-frequency behaviors (Cooper et al., 2007). Target behaviors included academic engagement/motivation as well as disruptive behavior. Each target behavior had identical
definitions as the DBR definitions. For example, Disruptive Behavior included criteria such as “calls out,” and Academic Engagement/Motivation included criteria such as “finishes work on time.” A student will meet criteria for displaying the target behavior if they demonstrate the behavior at least once during the 15-second time period.

Procedure

Rating Design and Procedure

Graduate student observers completed their DBRs for academic engagement or disruptive behavior via Qualtrics. Specifically, observers were presented with a screenshot of a 12–minute video clip with a student identified by a red circle, then were asked to watch the video clip of that student engaged in a math lesson and provide a DBR on a specified target behavior (either academic engagement or disruptive behavior). The Qualtrics survey was set up to enable observers to work at their own pace (e.g., complete a portion of the survey, save their progress, and return at a later time to continue). However, observers were instructed not to take breaks while watching a video clip and to complete their ratings within one week in an effort to standardize the amount of time each observer spent on the task. The majority of observers finished their ratings within three days (80%), while one participant took five days and one participant took eight days to complete the survey.
Independent Variable

In the current study, graduate student observers were randomly assigned to one of two conditions: a) a rater training group ($N = 4$) and b) a control group ($N = 6$). Originally, 11 graduate student observers were recruited, but one participant from the training group dropped out of the study, resulting in the unequal group numbers. However, chi-square tests of independence were conducted after randomization to ensure the two groups were not significantly different from one another on meaningful demographic variables related to prior experience with DBRs, $x^2(3) = 3.8889$, $p = .27$, and year in their program of study, $x^2(4) = 3.75$, $p = .44$. Next, the training group received training on the DBRs, which was informed by an existing DBR training module created at the University of Connecticut (University of Connecticut, n.d.) and included training components thought to be most effective at reducing rater effects, such as providing detailed behavior definitions and opportunities to practice with feedback (Briesch et al., 2010; Chafouleas, 2011; Chafouleas, Christ, & Riley-Tillman, 2009; Christ et al., 2010; Volpe & Briesch, 2012; Wickerd & Hulac, 2017). Because using a behavior rating tool appropriately is considered a learned skill, the training module was also informed by behavioral skills training, which is commonly viewed as best practice methods for teaching new skills. Behavioral skills training has four main components: instruction, modeling, role play, and feedback. Additionally, prior research on DBRs with training also include these elements (Chafouleas et al., 2010). As such, the training module for the current study included these components.
Specifically, the 30-minute training module was provided through a Qualtrics survey and included the following three primary components: 1) **definitions** for each behavioral indicator of disruptive behavior (e.g., definitions of “noisy,” “calls out,” etc.) and academically engaged behavior (e.g., definitions of “stays on task,” “starts tasks promptly,” etc.), 2) video clips of students in a classroom with a **demonstration** of ratings and explanations for each target behavior, and 3) an **opportunity to practice** rating a set of four additional video clips of students with feedback and explanations for the researcher-assigned ratings. The practice training videos allowed observers to make three attempts at assigning the “correct” rating until the observers’ ratings were within one category of the expert rating, after which the expert ratings were revealed. The expert ratings were obtained after the researchers first independently rated each clip, then discussed any discrepancies and came to a consensus. Allowing observers multiple attempts to rate within one point of the expert ratings also served as an integrity check to ensure that observers in the training condition adequately attended to the training module. Results from the training module indicated that raters typically needed more attempts (~2-3 attempts) to get within one category of expert ratings for Academic Behavior as they did for Disruptive Behavior (~1-2 attempts). Finally, observers in the training condition were asked to provide qualitative feedback on the helpfulness of the DBR training module at the end of the study.

Control group observers received the standard DBR-MIS form used by Volpe and Briesch (2012) for each target behavior that contains general DBR completion instructions and a list of the target behaviors. No other guidelines were provided.

Finally, observers were instructed to complete the DBR ratings independently to prevent collusion and diffusion of the intervention effects. All observers also had the opportunity to
indicate whether or not they felt unsure about a particular rating, followed by a prompt to indicate the reason for feeling unsure (e.g., “It wasn’t clear which value to assign,” “The item was irrelevant to the situation,” “Other [explain]”) as a means to gather additional qualitative data regarding how each observer approached each DBR item.

**Dependent Variables**

Rater severity/leniency was estimated using the MFRM model and SDOs were used as a “gold standard” measure of student disruptive and academic engagement/motivation behavior. Model-data fit was examined in multiple ways. First, overall model-data fit was evaluated by examining the standardized residuals. Good model-data fit was determined if < 5% of standardized residuals were ≥ |2| and < 1% of standardized residuals were ≥ |3| (Linacre, 2018). Second, mean-square (MSQ) infit and outfit statistics were examined to determine the degree to which measurement facet elements (i.e., individual students, items, and raters) fit the Rasch model. MSQ infit/outfit statistics identify measurement facet elements that either misfit (MSQ infit/outfit values ≥ 2) or overfit (MSQ infit/outfit values < .50) the MFRM model (Linacre, 2002).

Differences in rater severity/leniency across raters were measured by the $Q$ statistic, an omnibus significance test for rater effects. A significant $Q$ statistic indicates that individual raters demonstrate significantly different levels of severity/leniency in their ratings from each other. In addition, differences in rater severity/leniency between raters in the two training conditions differential were evaluated through an analysis of differential group functioning.
Finally, relations between raw DBR ratings, Rasch-calibrated DBR ratings, and SDO ratings were examined in order to determine the degree to which rater effects reduce the validity of behavior assessment conducted with DBRs. SDO ratings have been traditionally used as the “gold standard” method for assessing student behavior (Briesch et al., 2010; Chafouleas, Jaffery, et al., 2013; Chafouleas et al., 2005; Riley-Tillman et al., 2008; Riley-Tillman et al., 2009) and so stronger relations with SDO ratings were considered evidence of stronger external criterion validity. Two undergraduate students served as SDO raters. The students were provided SDO training with a set of six practice video clips, reaching 90% interobserver agreement using unweighted kappa prior to rating videos for the study. Raters then rated the student videos using the partial interval recording SDO method and interobserver agreement was evaluated after each video. When interobserver agreement fell below 90%, the raters were retrained and then rated the video again. For these SDOs, disruptive behavior was marked as “observed” between 0% to 16.5% of intervals during observations. In other words, the target behavior was observed for 0% of intervals in some observations and up to 30% of intervals in other observations. For SDOs of academic engagement, the target behavior was observed between 21.667% and 73.25% of intervals.
Data Analysis Plan

Question 1: Does rater training significantly reduce rater effects when compared to no rater training?

Mean-square (MSQ) infit and outfit statistics were examined to determine model-data fit. MSQ infit and outfit statistics that fall below 0.5 indicate that the data function in a dichotomous fashion like a “lightswitch,” such that the measurements tell us if a student performs a behavior at a high/low level rather than at what level of proficiency the student is able to perform that behavior. MSQ infit and outfit statistics that fall at or above 2.0 indicate that at least 50% of the variance in the data is due to unmodeled “noise” rather than true differences in behavioral skill level. In addition to MSQ infit and outfit statistics, standardized residuals were examined to determine model-data fit. Good model-data fit is indicated when 1% or fewer of residuals are > |3| and 5% or fewer of residuals are > |2|. Once good model-data fit is established, Research Question 1 will be answered by inspecting evidence of differential group functioning to determine if observer severity/leniency operates as a function of training group membership.

Question 2: Do rater effects reduce the validity of DBR scores?

Question 2 was answered by comparing the relation between the observer-assigned DBR ratings and SDO ratings with the relation between the Rasch-calibrated DBR ratings and SDO ratings. Rasch-calibrated DBR ratings are referred to as “fair average” scores and these scores represent the average score that a student would be assigned from a rater with average severity/leniency on items with average difficulty. Reduced validity would be evidenced by a
weaker relation between the observer-assigned DBR ratings and SDO ratings when compared to
the relation between Rasch-calibrated DBR ratings and SDO ratings.
CHAPTER IV

RESULTS

Question 1: Does rater training significantly reduce rater effects when compared to no rater training?

Academic Engagement

Graduate student observer ratings ranged from 0 (Never) to 6 (Always) across all observations. In other words, raters indicated ratings of Never (i.e., rating 0) for some observations up to Always (i.e., rating 6) for other observations. As a whole, observers demonstrated a tendency to assign high ratings (i.e., Often, Very Often, or Always) on Academic Engagement items. In particular, “Finishes work on time,” “Actively participates in class,” “Stays on task,” and “Starts tasks promptly” ratings were all negatively skewed, whereas ratings on “Works independently” were approximately normally distributed. This distribution of ratings was somewhat expected, given that it has also been reported in other studies using DBRs to measure academic engagement (Anthony et al., accepted).

Model-data fit was determined to be acceptable by an inspection of standardized residuals, mean-square (MSQ) infit and outfit statistics for the rater and item facets, and
category usage statistics to ensure that the 7-point rating scale functioned as intended. Approximately 1.45% of standardized residuals were $> |3|$ and 4.64% of standardized residuals were $> |2|$. Moreover, MSQ infit and outfit statistics for the rater facet ranged between .53 and 1.54 and MSQ infit and outfit statistics for the item facet ranged between .67 and 1.84. Finally, average student measures on Academic Engagement increased monotonically with rating categories, which indicates that higher ratings corresponded to higher levels of academic engagement.

Altogether, observers’ demonstrated significantly different levels of rating severity/leniency when compared to one another, $\chi^2(9) = 66.1, p < .001$, and these observer rater severity/leniency measures were reliably different (separation reliability = .88). However, training did not explain these differences on Academic Engagement items, $\chi^2(1) = 0.00, p = .89$, and the spread of rater severity/leniency measures was similar for observers in the two rater training conditions: 1.41 logits wide for observers in the rater training condition and 1.46 logits wide for observers in the control condition. Moreover, there was little to no evidence of differential rater functioning across the two observer rater training conditions for either items, $\chi^2(10) = 9.9, p = 0.45$, or students, $\chi^2(14) = 6.3, p = .96$. The only exception was noted when making a closer inspection for potential bias at the individual element level for both items and students, which revealed that graduate student observers in the training group were significantly less severe (i.e., assigned higher ratings) than observers in the control group when rating students on the item “Actively participates in class,” $t(57) = -2.74, p = .008$, but there was no evidence of differential rater functioning for any individual student.
Across all observations, ratings for Disruptive Behavior ranged from 0 (*Never*) to 6 (*Always*). However, graduate student observers demonstrated a tendency to assign low ratings (i.e., *Never*, *Rarely*, or *Sometimes*) on all Disruptive Behavior items, resulting in a positively skewed ratings distribution. Notably, some empty rating categories were observed wherein no observer assigned any student a particular rating in the entire dataset. Specifically, no observers assigned ratings of *Sometimes* (3), *Often* (4), *Very often* (5), or *Always* (6) on “Out of seat/area” for any student and no observer assigned ratings of *Always* on either “Talks to classmates when inappropriate” or “Calls out” for any student. This pattern of ratings was also somewhat expected, given that similar findings have been reported in prior research with DBRs for disruptive behavior (Anthony et al., accepted). As a result, a “dummy” student with a low weight (.0001) was added to the dataset in order to estimate rater effects for the entire scale ranging from *Never* (0) to *Always* (6) as suggested by Linacre (1989).

Nevertheless, model-data fit did not meet conventional standards for Disruptive Behavior items, approximately 8% of standardized residuals > |3| and 2.67% of standardized residuals > |2|. Moreover, MSQ infit and outfit statistics ranged between .26 and 3.56 for the rater facet and between .62 and 3.90 for the item facet, indicating the presence of unexpected ratings (MSQ infit and outfit values ≥ 2.0). Residuals were further inspected in order to identify what raters and/or items may have contributed to the poor model-data fit. This inspection revealed that two raters demonstrated MSQ infit and outfit values > 2 and one item (“Out of seat/area”) demonstrated MSQ infit and outfit values > 2. Moreover, the majority of unexpected ratings (over two-thirds)
occurred on the problematic “Out of seat/area” item. In all cases, graduate student observers assigned higher ratings on this item than was expected by the MFRM model. Consequently, the “Out of seat/area” item was removed and the reduced dataset was reanalyzed without this item.

With the removal of “Out of seat/area,” 4.44% of standardized residuals were > |3| and 1.11% of standardized residuals were > |2|. Moreover, MSQ infit and outfit statistics ranged between .28 and 1.91 for the rater facet and between .77 and 1.45 for the item facet; average measures increased monotonically, with categories indicating that higher ratings corresponded to higher levels of disruptive behavior as intended. As a result, good model-data fit was achieved.

Altogether, graduate student observers demonstrated significantly different levels of rater severity/leniency, $\chi^2(9) = 98.9, p < .001$, and these observer rater severity/leniency measures were reliably different (separation reliability = .94). Moreover, differences in graduate student observers’ rating tendencies towards severity/leniency could be attributed to the rater training condition to which they were assigned, $\chi^2(1) = 33.6, p < .001$, and these observer rater severity/leniency differences between groups were also reliably different (separation reliability = .97). Specifically, observers in the rater training group were significantly less severe (i.e., assigned higher ratings) when compared to observers in the control group. Moreover, the spread of rater severity/leniency measures was .83 logits wide for observers in the rater training condition and 2.31 logits wide for observers in the control condition. This indicates that the range of rater severity/leniency measures was nearly three times as large for the control group when compared to the rater training group.

There was also some evidence of rater bias on student measures. Although as a group there was no evidence of bias, $\chi^2(16) = 9.9, p = 0.87$, upon further inspection of individual
student measures, there was evidence of bias by rater training condition on ratings for one particular student, such that the graduate student observers in the training condition were significantly less severe (i.e., assigned higher ratings) as a group when compared to graduate student observers in the control group, $t(35) = -2.28, p = 0.03$. Further inspection indicated that this result may have been influenced by two observers, one from each training group condition, who were especially different in their tendencies towards severity/leniency than observers in the opposite condition.

There was also some evidence of rater bias on item measures, $X^2(8) = 16.8, p = .03$. There were two Disruptive Behavior items that accounted for the majority of the bias: “Noisy” and “Talks to classmates when inappropriate.” Graduate student observers in the rater training condition were significantly less severe (i.e., assigned higher ratings) as a group when compared to observers in the control group when rating students on “Noisy,” $t(63) = -3.18, p = .002$, and graduate student observers in the rater training condition were significant more severe (i.e., assigned lower ratings) as a group when compared to observers in the control group when rating students on “Talks to classmates when inappropriate,” $t(61) = 2.50, p = .02$. Moreover, pairwise differences in graduate student observers from each rater training condition were evenly distributed on items and did not appear to be influenced by a select subgroup of observers with “extreme” rating tendencies.
Question 2: Do rater effects reduce the validity of DBR scores?

Table 1 contains the validity coefficients between SDO ratings and DBRs. Overall, high, statistically significant, positive correlations were noted between SDO ratings and DBRs regardless of training status ($r = .906-992$). This indicates good validity for DBRs in general as behavior measurement tools. More specifically, neither rater training nor many-facet Rasch measurement calibration to account for rater effects improved the validity of DBRs of academic engagement. However, rater training significantly improved the validity of DBRs for Disruptive Behavior when compared to both ratings obtained from the control group (95% CI: .035, .343) and many-facet Rasch measurement-calibrated ratings controlling for rater effects (95% CI: .023, .234). Moreover, although many-facet Rasch measurement-calibrated DBRs were more strongly related to SDOs than DBRs from graduate student observers in the control condition, this difference was not statistically significant.

Table 1. Pearson Correlations Between DBRs and SDOs Disaggregated by Training Status

<table>
<thead>
<tr>
<th>Average SDOs</th>
<th>Average DBRs</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Control</td>
<td>Fair Average</td>
<td>Significance</td>
<td></td>
</tr>
<tr>
<td>Academic Engagement</td>
<td>.946</td>
<td>.920</td>
<td>.929</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disruptive Behavior</td>
<td>.992</td>
<td>.906</td>
<td>.938</td>
<td>T &gt; C, FA</td>
<td></td>
</tr>
</tbody>
</table>

Note. Average SDOs = average percent of intervals each behavior was rated to have occurred from systematic direct observations; Average DBRs = observed average direct behavior ratings across all raters in each group on all items for a particular student; Fair Average = observed average direct behavior ratings across all raters in each group adjusted to account for observer rater severity/leniency and item difficulty; T = training condition; C = control condition; FA = fair average.
CHAPTER V

DISCUSSION

Students’ levels of academic engagement and disruptive behavior are predictors of lifelong success (Jimerson et al., 2000; Kretschmer et al., 2014; Raffaele Mendez, 2003; Thompson et al., 2011). Therefore, it is imperative that teachers have efficient, feasible, and accurate tools to identify students who may need additional behavioral support. DBRs offer a potential solution, but prior studies have not yet addressed the best training method to reduce well-known between-rater rater effects. The primary purpose of this study was to determine the impact of rater training compared to no rater training on between-rater rater effects. A secondary purpose of this study was to determine the degree to which rater effects impact DBR validity. Our study took a comparative approach, investigating the differences in rater effects between a group of raters who were randomly assigned to complete a DBR training module and a group of raters who were randomly assigned to receive no training.

Results demonstrated that for Academic Engagement, raters were significantly different from one another in their tendencies towards severity/leniency, but these differences were not explained by rater training. For Disruptive Behavior, however, training group status explained differences in rater severity/leniency such that participants in the training group were significantly less severe in their ratings than those in the no training group. In regards to validity, results demonstrated that regardless of training, DBRs had good relations with SDOs. This
indicates good validity for DBRs as a measurement tool. For Disruptive Behavior, specifically, relations between DBRs and SDOs were significantly stronger for the training group compared to the no training group, indicating better validity was achieved on this target behavior when graduate student observers received training. Additionally, Rasch calibration to account for rater severity/leniency and item difficulty did not significantly improve the validity of ratings for Academic Engagement or Disruptive Behavior.

Impact of Training on DBRs

Results of the current study demonstrated that rater training reduced rater effects for Disruptive Behavior but did not reduce rater effects for Academic Engagement. One possible explanation for this is that there were fewer opportunities to observe disruptive behavior than academic engagement in the student classroom videos, making differences in rater effects much more likely. While graduate student observers used the entire scale when rating students on academic engagement, the upper half of the scale was not used by a single rater on Disruptive Behavior items. Because there were no students displaying extreme levels of disruptive behavior, it may have been more difficult to discriminate between students’ lower levels of disruptive behavior, allowing for rater effects to more strongly influence ratings on these items. Thus, rater training was impactful for these items. Additionally, it is possible that training did not influence rater effects for academic engagement because all graduate student observers already had similar ideas of what high and low levels of Academic Engagement should “look like” in a classroom. This was supported through the rater severity/leniency measures, which indicated that the two
groups had a similar spread for Academic Engagement, whereas the spread of the rater severity/leniency measures for the control group was twice as large as that of the training group for Disruptive Behavior. Results from rater training seem to support this idea – graduate student observers in the training condition took more attempts (typically three) to get within 1 point of expert ratings on Academic Engagement, whereas they took between one or two attempts to get within 1 point of expert ratings on Disruptive Behavior. This may indicate that it was easier for observers to calibrate their own ratings with expert ratings for disruptive behavior than it was to calibrate their ratings for academic engagement, possibly due to more preconceived notions about displays of academic engagement compared to displays of disruptive behavior.

Qualitative data regarding the most helpful and least helpful aspects of training were also collected from observers in the rater training condition. Graduate student observers reported that opportunities for practice were most helpful, while demonstrations of ratings were least helpful. Common suggestions for training improvement included additional opportunities to practice with feedback, as well as viewable definitions for the entirety of the rating process.

Validity of DBRs with Rater Training

Overall, validity coefficients between DBRs and SDOs were in the high range regardless of training status. However, due to individual differences in rater severity/leniency, it is clear that student scores can be highly impacted based on who completes the DBR. There was significant spread of rater/severity for both Academic Engagement and Disruptive Behavior items, meaning that a student’s score could be an overestimate or an underestimate of their true behavioral skills.
proficiency based on the tendencies of the observer. This would be especially true in cases where a student is rated by only one observer and therefore receives only one rating. Additionally, the use of upper rating categories was extremely sparse for Disruptive Behavior. Therefore, in contrast to conventional thinking from CTT reliability studies (Simms et al., 2019), providing additional rating categories may not be best practice. It may be more beneficial to provide fewer rating categories so that observers are encouraged to use the entirety of the rating scale.

Finally, our study provides implications for applied practice. Current practitioners rely heavily on rater-mediated assessments to identify students who may be in need of further behavior support. Specifically, a recent nationwide survey revealed that 97% of school psychologists reported using BRSs regularly, and 20% reported using BRSs as screening tools (Benson et al., 2019). Additionally, of participants who use BRSs as screening tools, 35% reported using a single screening tool in isolation for decision-making purposes, and 23% reported using two or three BRSs for decision-making purposes (Benson et al., 2019). This demonstrates that rater-mediated assessments carry heavy weight when considering students for whom to allocate resources. Therefore, when using DBRs for screening purposes, schools ought to strive to obtain the highest levels of validity possible in order to ensure fair distribution of behavioral support. In the current study, the observed average scores without Rasch calibration were highly correlated with DBRs, and it therefore may be advisable to use multiple raters for the same student and simply use the observed average score. However, our results indicate that training raters to use DBRs produces the highest levels of validity when compared to no training or using Rasch calibration. As such, it may be advisable for practitioners to consider training
raters on DBRs specifically for disruptive behavior, as this produced the highest validity coefficients.

Conclusion

In conclusion, DBRs have overall high concurrent validity with SDOs and are therefore useful tools for measuring behavior in the school setting. DBR training is most useful for reducing rater effects on Disruptive Behavior items and did not substantially impact ratings of Academic Engagement items. DBR training for disruptive behavior produces significantly higher validity coefficients, indicating that DBR training may be necessary for ratings of disruptive behavior to be interpreted as valid indicators of students’ classroom behavior. Further, training is recommended for DBRs of academic engagement in order to obtain the highest levels of validity possible.

Additionally, practitioners may consider using DBRs in the school setting with multiple raters for a single student, taking the average of the scores in order to produce an observed average score. While not statistically significant, Rasch calibration did improve the validity of ratings for academic engagement and disruptive behavior in this study, and practitioners ought to consider using Rasch calibration in order to achieve these higher levels of validity for high-stakes decisions regarding behavior support. Additionally, MFRM revealed unexpected ratings on Item 3 for Disruptive Behavior, which caused poor model-data fit. Without MFRM, practitioners would be unaware of problematic items that may cause unexpected ratings for students, influencing the allocation of student support services. Therefore, in addition to training,
the use of MFRM for rater-mediated assessments is recommended in order to obtain appropriate ratings for decision-making purposes.

The empty rating categories for Disruptive Behavior on Items 3, 4, and 5 also have implications for future studies. Prior research on CTT indicates that additional points on an ordinal scale may be optimal (Simms et al., 2019), but results of the current study demonstrate that observers did not use the upper half of the rating scale (i.e., Sometimes, Often, Very often, Always) for three items on Disruptive Behavior. Because CTT does not require the investigation of category usage statistics, it is possible that the problem of empty rating categories has always been present but never addressed. MFRM requires the usage of the entire rating scale in order for the proper functioning of the model, revealing any empty rating categories. Future studies ought to investigate category usage with fewer rating categories in an attempt to identify the optimal number of rating categories for complete scale usage.

Feedback from graduate student observers in the rater training group suggests that additional opportunities to practice might be a beneficial component to future rater training modules, as well as access to DBR definitions throughout the rating process. Limitations of this study include the nature of the virtual observations as compared to more realistic in vivo observations, as well as the fixed range of behavior displayed by the students in the videos. In vivo observations may allow for a wider range of behavior to be observed. Additionally, the videos used in the current study contained students in a middle-school setting, and results may not be entirely generalizable to other academic settings. Future studies ought to consider the impact of training for ratings of students in upper and lower grade levels. Another limitation includes the use of graduate students rather than school personnel as observers. Additionally, the
graduate student observers in this study were primarily White, female students from the same
graduate program, and future studies ought to consider including a wider range of participants to
further investigate differences in rater tendencies. Future studies should address these limitations,
as well as investigate the impact of providing DBR definitions throughout the observation on
rater effects.
REFERENCES


University of Connecticut. (n.d.). *Direct behavior rating training site.* https://dbrtraining.education.uconn.edu/


APPENDIX A

DIRECT BEHAVIOR RATINGS
Thank you for your participation in this study! Please enter your participant ID found in the email that you received with the link to this survey.

The following survey will take approximately 2.5 to 3 hours to complete. If at any point you need to stop and take a break, your answers will be saved and you can continue when you come back. However, please do not take a break in the middle of viewing a video or completing the associated rating. Please do not spend longer than 20 minutes viewing a video and completing the associated rating.

In this survey, you will be asked to watch 10 minute video clips of students in a middle school classroom. Prior to each video, there will be a screenshot that indicates which student you will observe. Please watch the identified student for the entirety of the video, paying close attention to the target behavior (either Academic Engagement or Disruptive Behavior).

After viewing the video, you will complete a Direct Behavior Rating (DBR) on the identified student. Each DBR contains 5 items to complete related to the target behavior.

There will be a total of 15 videos to view, with 15 associated DBRs to complete. Please maintain attention to the task to the best of your ability throughout the survey. If you feel yourself becoming fatigued, finish the video you are currently watching and the associated DBR, and take a break.
Target Behavior: Disruptive Behavior

Below is a list of behaviors that students may demonstrate in the classroom. Please read each item and rate how the child behaved during the observation period.

<table>
<thead>
<tr>
<th>Behavior</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Never</td>
</tr>
<tr>
<td>Clowns around</td>
<td>○</td>
</tr>
<tr>
<td>Noisy</td>
<td>○</td>
</tr>
<tr>
<td>Out of seat/area</td>
<td>○</td>
</tr>
<tr>
<td>Talks to classmates when inappropriate</td>
<td>○</td>
</tr>
<tr>
<td>Calls out</td>
<td>○</td>
</tr>
</tbody>
</table>
Target Behavior: Academic Engagement

Below is a list of behaviors that students may demonstrate in the classroom. Please read each item and rate how the child behaved during the observation period.

<table>
<thead>
<tr>
<th></th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Never</td>
</tr>
<tr>
<td>Works independently</td>
<td>●</td>
</tr>
<tr>
<td>Starts tasks promptly</td>
<td>●</td>
</tr>
<tr>
<td>Finishes work on time</td>
<td>●</td>
</tr>
<tr>
<td>Actively participates in class</td>
<td>●</td>
</tr>
<tr>
<td>Stays on task</td>
<td>●</td>
</tr>
</tbody>
</table>
Introduction to Direct Behavior Ratings

**What is a Direct Behavior Rating (DBR)?**

A DBR is a behavioral assessment instrument in which a target behavior (e.g., academic engagement, disruptive behavior, etc.) is identified and observed during a specified observation period, and then rated by the observer on some type of rating scale. It is thought to combine the benefits of both behavior rating scales and systematic direct observations by being both easy to complete and providing accurate information.

**What kinds of DBRs exist?**

There are two types of DBRs: single item scales (DBR-SIS) and multi-item scales (DBR-MIS). A DBR-SIS contains a single item for each target behavior (e.g., "rate the student's overall disruptive behavior"), while a DBR-MIS contains multiple items for each target behavior (e.g., "rate the students disruptive behavior on calling out, out of seat/area, clowning around, etc.").

Additionally, DBR-SISs typically assess the percent of time a student is observed to engage in a target behavior, whereas DBR-MISs typically assess the target behavior using a Likert-type rating scale.

**How are DBRs typically used?**

In school-based settings, DBRs have been used as both progress monitoring tools as well as universal screeners. Most of the time, schools use DBRs as screeners to identify students that may need more intensive behavior support in the classroom. DBRs have been developed to target
a number of behaviors, such as, academic engagement, disruptive behavior, respectful behavior, and compliance, as examples. Observations usually span a minimum of 10 minutes in duration and are typically conducted by a classroom teacher, paraprofessional, or other school personnel (e.g., school psychologist).

Target Behaviors Defined

The remaining components of this training will teach you how to administer DBR-MIS to assess two specific target behaviors: a) disruptive behavior and b) academic engagement.

In the sections below, we have defined these two target behaviors for you. Please review these definitions before proceeding with the training.

DBR Definitions

Disruptive Behavior

Clowns around is defined as the student making humorous sounds or gestures (e.g., cracking jokes, making silly faces, or joking around).

Noisy is defined as the student making sounds either verbally, by handling materials, or by performing other motor movements that cause noise that is noticeably louder than other noises in the room.
Out of seat/area is defined as observing the student away from their seat or desk/table area (e.g., walking around the room or standing up). This does not include a student who is standing at their desk with part of their buttocks or legs on their chair.

Talks to classmates when inappropriate is defined as the student talking to peers at a time that is not permitted (e.g., whole group lessons, independent work, etc.). This does not include instances in which students are permitted to work together.

Calls out is defined as the student talking out of turn (e.g., talking when the teacher is talking, talking without being called on by the teacher, or making unrelated comments to themselves, peers, or the teacher during the observation period). This does not include instances when the teacher asks for a choral response, or when the teacher allows students to shout out answers to questions.

Academic Engagement

Works independently is defined as maintaining focus and working on the assigned task without teacher support.

Starts tasks promptly is defined as the student beginning work on the current task within 10 seconds of the teacher prompt.

Finishes work on time is defined as the student completing the task by the end of the assigned time.

Actively participates in class is defined as the student giving full attention to the task as indicated by looking at the teacher when they are delivering instructions, responding when directed by the teacher to provide a response (e.g., raising their hand or providing a response
when called upon by the teacher), or on-topic conversations with peers during small group work.

**Stays on task** is defined as the student maintaining engagement. The student does not engage in off-task behavior, such as, looking around the room or leaving their seat/area.

Example Demonstration

**Academic Engagement**

**Works independently**

**Rating:** 4

**Explanation:** The student was rated a 4 for this item because although he did not receive teacher support, he also did not fully maintain focus on his task. He was also observed to turn around in his seat after not being called on by the teacher.

**Starts tasks promptly**

**Rating:** 5

**Explanation:** The student was rated a 5 on this item because he immediately began writing after the class was given a teacher directive the majority of the time in the video clip. The student would have been assigned a higher rating if he had started working promptly after every teacher directive.

**Finishes work on time**
Rating: 4

Explanation: The student was rated a 4 on this item because he appeared to complete most tasks on time. However, he did not complete all of the tasks on time.

Actively participates in class

Rating: 5

Explanation: The student was rated a 5 on this item because he engaged and kept his hand raised for the majority of the lesson. The student was not rated higher on this item because he was also observed to turn around in his seat and look at the back of the classroom at times in the video clip.

Stays on task

Rating: 4

Explanation: The student was rated a 4 on this item because he demonstrated fairly good attention to the task, at the same time, he also engaged in a few instances of off-task behavior, like turning around in his seat or looking on the floor for his pencil.

Disruptive Behavior

Clowns around

Rating: 1

Explanation: The student was rated a 1 on this item because although he was not loudly making
jokes to the class, he and the student seated next to him were laughing quietly together towards the end of the video. This elevates his rating from a 0 to a 1.

Noisy

Rating: 4

Explanation: The student was rated a 4 on this item because he was tapping on his desk, which was loud enough to be heard on the video. He did this multiple times throughout the video, but each occurrence lasted only a few seconds. For this reason, he was rated a 4.

Out of seat/area

Rating: 0

Explanation: The student was rated a 0 on this item because he did not leave his seat/area during the video clip.

Talks to classmates when inappropriate

Rating: 4

Explanation: The student was rated a 4 on this item because he spent a portion of the video clip talking to the student seated next to him. This did not last the entirety of the video clip, which would have warranted a higher rating, but it did occur more often than not.

Calls out

Rating: 0
**Explanation:** The student was rated a 0 on this item because he did not call out during the video clip. Any comments that the student made were only audible to the student seated next to him and not loud enough to be captured on the video clip.

**Feedback Example for Practice Opportunities**

**Disruptive Behavior**

That's correct! You provided ratings within 1 category of the expert ratings. See below for a complete explanation:

**Clowns around:** 0

**Explanation:** The student was rated a 0 because he did not make any humorous faces, gestures, or sounds throughout the video clip.

**Noisy:** 2

**Explanation:** The student was rated a 2 because he was quiet for most of the video clip. However, he did cause noise when he got out his binder/folders and began handling them.

**Out of seat/area:** 0

**Explanation:** The student was rated a 0 because he did not leave his seat during the video clip.
Talks to classmates when inappropriate: 0

Explanation: The student was rated a 0 because although he looked at another student's paper, he did not talk to that student during the video clip.

Calls out: 0

Explanation: The student was rated a 0 because he did not call out during the video clip.

Academic Engagement

That's correct! You provided ratings within 1 category of the expert ratings. See below for a complete explanation:

Works independently: 4

Explanation: The student was rated a 4 because she completed tasks without adult redirection, but she did not maintain focus throughout the entire video clip.

Starts tasks promptly: 3

Explanation: The student was rated a 3 because there were a few instances where she was observed to write down the problem as soon as it was given by the teacher, but more often than not, she took a few seconds to begin writing.

Finishes work on time: 5
**Explanation:** The student was rated a 5 because she finished most of the problems given by the teacher promptly after she began working.

**Actively participates in class:** 4

**Explanation:** The student was rated a 4 because she raised her hand to answer a number of questions that the teacher asked, but she was not observed to participate consistently throughout the entire video clip.

**Stays on task:** 3

**Explanation:** The student was rated a 3 because she looked down at her hands in between problems, but she was observed to work on the teacher directed task during some of the video clip.