THE EFFECT OF AN INTENSIVE TEACHER TRAINING ON THE ACCURACY OF SOCIAL, EMOTIONAL, AND BEHAVIORAL SCREENING RESULTS

A Dissertation Presented to

the Faculty of the Graduate School

University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

Kristy Warmbold-Brann

University of Missouri-Columbia

Drs. Matthew K. Burns and Stephen Kilgus, Dissertation Supervisors

July 2017

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

THE EFFECT OF AN INTENSIVE TEACHER TRAINING ON THE ACCURACY OF SOCIAL, EMOTIONAL, AND BEHAVIORAL SCREENING RESULTS

Presented by Kristy Warmbold-Brann

a candidate for the degree of Doctor of Philosophy,

and herby certify that, their opinion, it is worthy of acceptance.

Matthew K. Burns Ph.D.
Department of Educational, School, and Counseling Psychology

Stephen P. Kilgus, Ph.D.
Department of Educational, School, and Counseling Psychology

Cheryl A. Offutt, Ph.D.
Department of Educational, School, and Counseling Psychology

Chad A. Rose, Ph.D.
Department of Special Education

ACKNOWLEDGEMENTS

The current project and my doctoral experience would not have been possible without so many people. I would like to recognize and thank the University of Missouri School Psychology faculty and graduate students for offering a remarkable training program and motivating me to strive for my best. I would like to thank my dissertation co-chairs, Drs. Matthew Burns and Stephen Kilgus. I cannot express my gratitude enough for your ongoing support and dedication to make the study possible. The project was left without an adviser and you both graciously offered your guidance without a second thought. As my adviser and research team leader, Dr. Burns also contributed greatly to my professional and personal growth and I greatly valued my time learning from him. Also, I would like to thank Dr. Melissa Maras for serving as a mentor over my five years at University of Missouri and fostering my interest in universal screening. Thank you for encouraging this project through countless screening conversations and applied work with schools. To my committee, I want to thank you for your willingness to discuss the project and think through logistical concerns and for always supporting the project. I very much enjoyed learning from you and the time spent discussing the project with you. Also, thank you to the Center for Social and Emotional Success for treating this study as your own and supporting the project. I know it was a lot of work and I sincerely appreciate everyone's time, especially Dr. Stephen Kilgus, Dr. Katie Eklund, Crystal Taylor, and Amanda Allen for countless hours coordinating and handling logistical concerns. A special thank you to Jared Izumi, Lisa Aguilar, Kayla Kilpatrick, Rosie O'Donnell, Regan Riley, Deija McLean, Jennifer Connelly, and Mike Van Wie for making the project possible through diligent observations.

Last but not least, I want to express my thanks to my incredible family and friends that filled my life with encouragement, laughter, and joy over these past five years. Thank you to my

parents and sister for always supporting my studies and never questioning my love of school and learning. I also want to thank my informal family of the Southbrook house for helping me stay sane and keep challenges in perspective. To a doctoral candidate's best friends, Ace and Brewer, thank you for always welcoming me home with a big smile and a wagging tail. My husband, Mark Brann, has been my greatest support during my graduate studies. Thank you for always making me laugh and encouraging my doctorate, even when it meant living apart and seeing me attached to my laptop. I cannot imagine these five years without you by my side serving as my cheerleader and informal advisor.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THE EFFECT OF AN INTENSIVE TEACHER TRAINING ON THE ACCURACY OF SOCIAL, EMOTIONAL, AND BEHAVIORAL SCREENING RESULTS

Kristy Warmbold-Brann

Drs. Matthew Burns and Stephen Kilgus, Dissertation Supervisors

ABSTRACT

The current study examined that effect of an intensive Frame of Reference teacher training on the accuracy of teacher-rated social, emotional, and behavioral screening results as compared to objective systematic direct observations of one to two students per classroom (n = 74). Teachers (n = 64) were randomized into an intensive Frame of Reference training or familiarization control condition. Results from multilevel analyses suggest no statistically significant improvement from an intensive teacher training on the difference scores of Academic Behavior and Disruptive Behavior. Nonparametric analyses were completed for Prosocial Behavior due to the lack of between-teacher difference and similarly found no improvement. The impact of the training on the Total Behavior results of all students (n = 1158) was also examined including the distribution, number of students identified per classroom, and Total Behavior Scores. It was hypothesized that the training would not impact Total Behavior Scores. The hypothesis was correct as the intensive training did not significantly alter the number of students identified, homogeneity of variance, or Total Behavior Scores. Limitations, directions for future research, and practical implications are reviewed in detail.

CHAPTER I: INTRODUCTION

Background

Schools are tasked with the challenge of not only teaching academics but also supporting the development of behavior and social-emotional health (Satcher, 2000). There is a critical need to support student social, emotional, and behavioral health because approximately 14-20% of students experience behavior and/or mental health difficulties (Kessler Rc, 2005) and about half of Americans will meet the criteria for a mental illness diagnosis at some point in their lifetime (O'Connell, Boat, & Warner, 2009). In addition, the onset of social, emotional, and behavior disorders is usually in childhood and adolescence, and these students are at an increased risk for dropping out of school and having educational difficulties (Breslau, Lane, Sampson, & Kessler, 2008).

Researchers have pinpointed a variety of risk factors that are connected to social, emotional, and behavioral disorders and it is believed that individuals have a higher likelihood of having a disorder when they are exposed to multiple risk factors (O'Connell et al., 2009). In addition, if children are not meeting developmental competencies, then they can be at-risk for social, emotional, and behavioral problems (Guerra & Bradshaw, 2008). Risk factors and developmental competencies can occur within the child or a variety of environmental contexts such as home, school, and community (Bronfenbrenner, 1977). It is crucial to find ways to identify and support these students in a proactive manner before major problems develop.

Statement of the Problem

School personnel can help prevent social-emotional difficulties with early intervention and universal screening for problems in order to identify students in need of supports (New Freedom Commission on Mental Health, 2003). Schools need a brief tool that provides an

indication of all students' social and emotional functioning to identify students for additional supports. Universal screening in schools often relies on teacher ratings of student behavior in universal screenings (Kamphaus & Reynolds, 2007), which involves teachers rating the perceived occurrence and frequency of behaviors as they occurred over the past 1 to 6 months (Merrell, 2001).

Universal screening for social, emotional, and behavioral (SEB) concerns requires reliable and valid tools that are brief and feasible while discriminating risk from a lack of risk (Glover & Albers, 2007). Furthermore, using a teacher report of social and emotional functioning assumes that teachers can be relatively accurate reporters and have an understanding of the behavior they are rating. Yet, research suggests that informants often differ from one another when rating the same student (Achenbach 1987; De Los Reyes et al., 2015) and teachers fail to identify self reported depressive symptoms (Cunningham & Suldo, 2014; Auger 2004). In addition to informant differences between different types of raters, generalizability research, where multiple raters rate the same student in the same context, found that a large proportion of the variance is often due to the rater (Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese, 2007; Briesch, Chafouleas, & Riley-Tillman, 2010; Minor., 2013). The large portion of the variance attributable to the rater suggests that raters interpret student behavior differently and may have different perceptions and attributions about the causes of the behavior (De Los Reyes & Kazdin, 2006). There is a need to reduce teacher differences in ratings and improve accuracy for SEB universal screening because the results of the ratings impact which students are identified for additional social and emotional supports.

Many sources encourage training teachers before they complete universal SEB screenings (Severson, Walker, Hope-Doolittle, Kratowill, Greshman 2007; Walker, 2010; Lane, 2012;

Weist et al., 2007). However, there is little research on the effects of training teachers for SEB screening and one study did not find supportive results (Moor et al., 2007). A related line of research studied the effect of several teacher training conditions on the direct behavior ratings (DBR) of targeted behavior on brief video clips, and found that training improved accuracy but the results were mixed on the ideal training format and level of intensity (Schlientz, Riley-Tillman, Briesch, Walcott, & Chafouleas; 2009; Harrison, Riley-Tillman, & Chafouleas, 2014; Chafouleas, Kilgus, Riley-Tillman, Jaffery, & Harrison, 2012; Chafouleas, Riley-Tillman, Jaffery, Miller, & Harrison, 2014; Chafouleas, McDougal, Riley-Tillman, Panahon, & Hilt, 2005; LeBel, Kilgus, Briesch, & Chafouleas, 2010). Some studies found success with a direct and more intensive rater training, called Frame of Reference training, for difficult to rate behavior (Chafouleas et al., 2012; Chafouleas et al., 2015). The DBR studies encourage the use of training for universal screening but results may not generalize because the ratings are brief and targeted rather than rating symptoms over several weeks in a naturalistic environment.

The organizational consultation literature also supports the use of training raters to increase accuracy and reduce measurement error (Hoyt & Kerns, 1999). Frame of Reference training is suggested as the ideal rater training format (Roch, Woehr, Mishra, & Kieszczynska, 2012), which creates a standard reference for raters to use and improves common understanding of the behavior and how to rate. Previous studies have yet to study Frame of Reference training for universal SEB screening in a naturalistic school environment. Frame of Reference training offers the potential to improve the accuracy of SEB teacher screening, which in turn will improve SEB identification and intervention service delivery.

Purpose of the Current Study

Mental health screening often includes universal teacher ratings of behavior to identify students in need of additional supports. However, previous research suggests that teachers may rate students differently based on their perceptions, which could lead to inaccurate results in rating students. If teachers are not rating accurately, then students may be falsely identified for additional supports or not be identified when there are social and emotional concerns and intervention is needed. Misidentifying students causes schools to waste resources or miss an opportunity for early intervention. Universal SEB screening should assist schools in appropriately allocating resources but this relies on relatively accurate teacher ratings. To improve accuracy of screening results, training teachers using an established rater training method may offer a solution. Frame of Reference training could improve accuracy but has yet to be studied in the context of universal teacher SEB screening in a school setting. This study builds on the DBR research by examining Frame of Reference teacher training to universal SEB screening of behavior in a naturalistic setting. The present study aims to examine the effect of training teachers with FOR training on the accuracy of SEB screening results in comparison to direct observations of behavior. Thus, the study will seek to answer the following research questions:

Research Questions:

1. What is the effect of an intensive teacher training on the relationship between teacher ratings on universal screenings and independent systematic direct observations for academic behavior?
2. What is the effect of an intensive teacher training on the relationship between teacher ratings on universal screenings and independent systematic direct observations for social

behavior, including prosocial and disruptive behavior?

3. What is the effect of an intensive teacher training on the distribution, number of students identified, and scores for SAEBRS teacher ratings of Total Behavior?

Given the review of research, it is hypothesized that the intensive teacher training condition will lead to lower differences between the SAEBRS and systematic direct observations of behavior. Additionally, the training is not expected to impact the distribution, number of students identified, or Total Behavior ratings because that was not the focus of the training.

Definition of Key Terms

Social, Emotional, and Behavioral (SEB): SEB includes social, emotional, and behavioral factors, which will each be defined. Social characteristics refers to a student's ability and actions to form and maintain age-appropriate relationships while emotional factors correspond to a student's ability to regulate emotions, adapt to change, and cope with stressful situations (Kilgus, Eklund, von der Embse, Taylor, & Sims, 2016). Lastly, behavioral factors allude to aggressive behavior, inattention, and impulsivity (Achenbach et al., 1987).

Frame of Reference (FOR) training: A method of training raters to develop a common heuristic and perception of performance and frequency of behavior (Woehr, 1994).

Systematic Direct Observation (SDO): A direct measure of student behavior that includes a priori determined operational definitions of specific behaviors and scoring of behaviors in a standardized manner (Hintze, Volpe, & Shapiro, 2008).

Universal Screening: A method to identify students in need of early intervention, often applied to academic or SEB difficulties in schools. Universal screening involves briefly assessing all students to determine risk status (Glover & Albers, 2007).

Social, Academic, and Emotional Behavior Risk Screening Scale (SAEBRS): A teacher behavior rating screening tool to identify students at risk for social skill, emotional regulation, and academic readiness concerns (Kilgus, Sims, von der Embse, & Taylor, 2015).

Assumptions

There are several underlying assumptions regarding this study. First, it is assumed that the sample of teachers and students will be representative of the greater population. Additionally, one makes the assumption that the SEB measures reflect the SEB constructs they are intended to measure. Next, the study depends on the ability of independent observations of social and academic behavior to provide greater estimate of "true" behavior compared to the brief teacher screening. Lastly, the study assumes that teachers are independent raters and that the correlation between the teacher ratings and observations will accurately represent the relationship between the measures.

Delimitations

For this study, there are three delimitations that predetermine the boundaries and limit the study. First, only two students near the risk cut-off were chosen from each classroom, which limits the generalizability of the study. Additionally, the study measured student on-task, disruptive, and prosocial behavior with one to four 15-minute observations. The observation time may not be enough time to accurately capture measurements of student academic on-task, disruptive, and prosocial behavior. A third delimitation is that only students in kindergarten through 5th grade and two 8th grade students were observed, which limits the results for academic and social behavior to elementary school.

CHAPTER II: LITERATURE REVIEW

This chapter reviews relevant literature for the study's purpose. First, SEB screening systems and behavioral assessment will be reviewed. Then, research regarding concerns with rater accuracy will be examined. Lastly, literature for reducing accuracy concerns through training will be reviewed.

To assist schools in the prevention and early intervention of SEB difficulties, a public health tiered approach is recommended (National Association of School Psychologists, 2009; New Freedom Commission on Mental Health, 2003), which includes routine universal screening of the population to identify students in need of extra supports (Merrell, 2001). Glover and Albers (2007) offered considerations when schools select screening measures including contextual appropriateness, psychometric properties, and usability and feasibility. Mental health screening often includes teachers completing behavioral rating scales (Gresham & Elliott, 2008; Kamphaus & Reynolds, 2007), which has been shown to identify students at-risk for pervasive comorbid mental illness diagnoses (Essex et al., 2009). A screening system should offer relatively accurate results while maintaining feasibility to assist in decision-making efforts for selecting students for additional supports.

Behavioral Assessment

Child behavior is typically measured through rating scales or direct observation (Merrell, 2001). Rating scales involve informant(s) (i.e. parents, teachers, or the child) retrospectively judging the perceived occurrence and/or frequency of the behavior or emotion over the past 1 to 6 months (Merrell, 2001; Christ, Riley-Tillman & Chafouleas, 2009). Below I will discuss rating scales, systematic direct observation, and direct behavior ratings. I will define each, provide examples of each, and discuss factors that affect the validity of the resulting decisions.

Rating Scales

Rating scales are ideal for providing information on low frequency behavior over a period of several weeks and require a short amount of time to administer, score, and interpret but have limitations using informant perception and memory to provide an indirect measure of behavior and may not reflect behavior change over short periods of time (Merrell, 2001; Christ et al., 2009).

There are several types of rating scales that differ based on the purpose. First, there are broad rating scales which are ideal for universal screening and should be brief and feasible to complete and score on large numbers of students while covering a wide range of SEB constructs (Levitt, Saka, Romanelli, & Hoagwood, 2007). Broad scales should distinguish students at-risk from typically functioning peers (Levitt et al., 2007). Specialized instruments are a second form of rating scales that still cover a wide range of SEB concerns but in more detail than broad instruments and thus take longer to administer, score, and interpret (Levitt et al., 2007). Specialized instruments are ideal for selected or indicated prevention (Levitt et al., 2007). Lastly, targeted instruments are rating scales with a narrow focus such as just anxiety or depression and are best used for diagnosing or assessing the severity of symptoms (Levitt et al., 2007). Rating scale informants can be parents, teachers, or a self rating with a student rating their own symptoms (Smith, 2007). The ideal informant for rating scales depends on the type of behavior and the purpose of the assessment (Smith, 2007).

Validity evidence depends on the decision being made, format of the rating scale, informant, and the population (Levitt et al., 2007). Levitt and colleagues (2007) reviewed predictive validity research for broad, specialized, and targeted rating scales. Few broad rating scales were included in the review, but Goodman (2001) reported predictive validity for the

teacher rated Strengths and Difficulties Questionnaire (Goodman, 2001) in comparison to Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition diagnoses (sensitivity = .43, specificity = .95, negative predictive power = .94, positive predictive power = .44). For a targeted rating scale, the Behavioral Assessment System for Children (Reynolds & Kamphaus, 1998), self-ratings were compared to a diagnostic interview and higher sensitivity and lower specificity values were reported compared to the Strengths and Difficulties Questionnaire and positive and negative predictive validity was not reported (sensitivity = .64-.82, specificity = .51-.64; Doyle et al., 1997). There is a lack of research on broad instruments for use in SEB screening (Cook, Volpe, & Livanis, 2010). Additional threats to the accuracy of rating scales are discussed below under the Rating Accuracy Section. Best practices encourage school personnel to gather additional information about student need after the initial at-risk screening (Merrell, 2001; Severson et al., 2007).

Systematic Direct Observation

While rating scales are indirect measures of behavior, behavior observation involves directly observing behavior in the natural environment with an independent and objective observer (Merrell, 2001). To increase objectivity and standardization of observations, systematic direct observation (SDO) is recommended as best practice (Hintze et al., 2008). Systematic direct observation includes operational definitions of specific behaviors defined prior to observing and conducting the observation and scoring in a standardized manner (Hintze et al., 2008).

Methods of systematic direct observation include event, interval and time-sampling, and duration recording (Merrell, 2001). Event recording is indicating the number of times a behavior occurs during an observation period while interval recording is dividing the observation period

into intervals and determining if the behavior occurs at any point or during the entire interval. Momentary time sampling also involves intervals but records if a behavior occurs at specific times and duration recording notes the length of time a behavior occurs. Suen and Ary (1989) found that momentary time sampling was the most accurate and least biased time sampling method. Strengths of systematic direct observation include objectivity, directly recording behavior, and the inclusion of contextual information from the environment while weaknesses include time and resources as well as number of observations needed to gain a reliable and valid measurement of social and emotional behavior.

Researchers have examined the reliability of direct observations through generalizability theory. Hintze and Matthews (2004) found that 63% of the variance of observations of on-task/off-task behavior was attributable to individual student differences, while 24% of the variance was attributed to the residual, or error and reliability coefficients were low. Stichter and Riley-Tillman (2014) raised concerns regarding the validity of direct observations because coding schemes and procedures often differ based on the target behavior and the population, thus, specific tools used in practice often lack validity evidence and coding schemes may not be applicable to all populations. In an examination of seven published SDO codes, Volpe, Diperna, Hintze, and Shapiro (2005) reported that only three scales published convergent validity data but all seven scales provided published data on discriminant validity for identifying students with behavior problems from typically developing peers. For instance, the BOSS (Shapiro, 1996) demonstrated effect sizes from -.53 to 1.25 for direct observations of academic engagement and off-task behavior in predicting Attention-Deficit/Hyperactivity Disorder compared to a control group (DuPaul et al., 2004).

Direct Behavior Ratings

Direct Behavior Ratings (DBRs) are a third method for behavioral assessment that combine aspects of both rating scales and systematic observations. With DBRs, teachers complete a brief evaluative rating of targeted behavior at the time and place of the behavior. DBRs offer the efficiency of rating scales with the direct measure of behavior from observations and serve as in ideal method to monitor progress (Chafouleas, 2011). A disadvantage to DBRs is the need to rate several behaviors over a wide range of time to provide the same information as rating scales. Furthermore, DBRs lack the objective and independent rater of observations.

DBRs can vary based on the target words and number of items interpreted on the scale (Riley-Tillman & Chafouleas, 2009). Single-item scale DBR involves interpreting ratings for individual items (i.e. disruptive) while multiple-item scale DBR sums ratings across multiple items. The majority of DBR research has focused on single item scale DBR (Volpe & Briesch, 2012). Using generalizability theory, Volpe and Briesch (2012) examined the number of occasions needed to reach a dependable estimate of behavior at a reliability (.80) and found that results were improved for multiple-item scale compared to single-item. Acceptable reliability was obtained after 4 occasions for academic engagement/motivation multiple-item scale DBR, 12 occasions for disruptive behavior multiple-item DBR, and 17 occasions for academic engagement/motivation single-item scale, but reliability of .80 was not reached for disruptive behavior single-item scale after 100 occasions (Volpe & Briesch, 2012).

Validity for DBR depends on the scaling and target wording (Christ, Riley-Tillman, & Chafouleas, 2009). Christ, Riley-Tillman, Chafouleas, and Jaffery (2011) examined the criterion validity of different DBR single item scale in relation to SDO and found that validity coefficients were highest for academic engagement (.67) and disruptive behavior (.78) while other behaviors

fell below acceptable levels (.28-.37). In an investigation of the concurrent validity of DBRs as a screener, Chafouleas et al., (2013), found moderate significant correlations for elementary students between DBR items and the Student Risk Screening Scale (Lane, Parks, Kalberg, Carter, 2007; disruptive behavior = .69, academic engagement = -.64, respectful behavior = -.59) and the Behavioral and Emotional Screening System (Kamphaus & Reynolds, 2007; disruptive behavior = .63, academic engagement = -.70, respectful behavior= -.49).

Synthesis

Best practice for SEB screening includes administering rating scales for all children because rating scales take a short amount of time to administer and score while capturing functioning over a wide range of time and functioning (Levitt et al., 2007). After the initial screening, school personnel would follow up with direct observations to assist in classification and treatment planning and using DBRs to monitor progress (Merrell, 2001; Chafouleas, 2011). However, there has been a lack of validity research with broad rating scales that are used in universal SEB screening and the Strengths and Difficulties Questionnaire revealed low sensitivity and positive predictive power values, which is harmful to screening because at-risk students may not be identified (Glover & Albers, 2007). There is a lack of accuracy studies with rating scales and few studies of universal screening measures in practice. Research indicates that SDO and DBR can effectively distinguish students with SEB concerns, but the psychometric evidence varies based on the target behaviors and coding/rating scheme. SDO and DBR both lack standardization and require multiple administrations over several observation periods to provide reliable results, which may not be feasible for assessing the behavior of a wide range of students.

Rating Accuracy

Trusting the results of screening depends on the accuracy of the results. In the case of teacher ratings for SEB, the accuracy depends on the teachers completing brief universal risk assessments. Rater accuracy is the strength and direction of relationship between a rating and a standard "true" score of behavior to create an accuracy score or comparison (Sulsky & Balzer, 1988). Although validity and reliability are necessary for accuracy, they are often insufficient to judge rater accuracy (Sulsky & Balzer, 1998). Accuracy scores differ from validity because direct comparisons are made to a standard of behavior that serves as a "true" score (Sulsky & Balzer, 1998).

There are many factors that may impact rater accuracy. Cronbach, Gleser, Nanda, and Rajaratnam (1972) noted that rater accuracy was affected by observer characteristics and conditions of the observations as well as characteristics of the procedure and rating system. Cairns and Greene (1979) suggested several key assumptions of using raters including (a) that they have a common understanding of the trait (e.g., social skills), (b) have a common understanding of the behavior representing the trait (e.g. for social skills, the behavior of cooperating with peers), (c) have the ability and time to assess the frequency of occurrences, and (d) use a common scale for behaviors.

In a meta-analysis of bias in observer ratings, Hoyt and Kerns (1999) found that an average of 37% of the variance was due to the rater with sources coming from different interpretations of the rating scale and different appraisals of the same person. Examining the accuracy of social-emotional learning screening systems is imperative as the results impact if students are identified for additional supports within the school. There are several areas of concern with trusting the results of behavior rating scales that will be discussed below.

Informant Discrepancy

The cross-informant literature on behavior rating scales has extensively documented concerns with conflicting assessment of student behavior. When a parent or teacher rates a student's behavior, the results depend on the rater's familiarity with the student, the purpose for the assessment, goals for intervention, and context of working with the student and chances for observation (De Los Reyes & Kazdin, 2005). Therefore, discrepancies among raters are common. In a landmark study, Achenbach et al. (1987) completed a meta-analysis and found .64 inter-rater reliability between teacher, .28 between teachers and parents, and .20 inter-rater reliability between teacher and self-ratings. De Los Reyes et al. (2015) replicated Achenbach and colleague's findings by reviewing 341 studies published between 1989 and 2014 and reported similar low to moderate inter-rater reliabilities with a correlation of .28 between parents and teachers. While these studies highlight differences between informants, the three studies are limited in applicability to screening due to focus on full length assessments rather than brief and broad universal screening systems. Informant discrepancies raise concerns with assessment, classification, and treatment of child problems, which could apply to screening systems (De Los Reyes & Kazdin, 2006).

When considering informant discrepancies and rater accuracy, one must take into account the type of behavior being rated as several studies point to differences in behaviors with teachers demonstrating less accuracy and reliability with identifying internalizing symptoms. Reliability coefficients differ more for internalizing symptoms compared to externalizing (Kolko & Kazdin, 1993; Mattison, Bagnato, Mayes, & Felix,1990; McConaughy, Mattison, & Peterson, 1994). Auger (2004) examined teacher accuracy in rating middle school student depressive symptoms and found a low relationship when compared to student self-ratings (r = .22) and only 19% of the

judgments for the five identified depressed students were correct. Selection bias was a potential concern with this study as only 37% of students returned parental consent and a large portion of the sample, 13%, received special education services. In another study, teachers nominated students displaying depressive and anxiety symptoms and the authors compared the nominations to student self-ratings of internalizing behaviors and found that teachers fail to identify depression (sensitivity = .50) and anxiety (sensitivity = .41) in children about half of the time and have some false identifications (specificity = .16 and .18; Cunningham & Suldo, 2014). Both Auger (2004) and Cunningham and Suldo (2014) were limited by the assumption that the self-ratings provided an accurate measure of internalizing behavior and there were no reliability and validity evidence for the teacher identification method because the one used was teacher nomination (Cunningham & Suldo, 2014). Moreover, Auger (2004) created a rating scale for the study and did not present psychometric evidence for the new scale.

Hinshaw, Han, Erhardt, and Huber (1992) studied the ability of parent and teacher rating scales to predict student internalizing and externalizing behavior through observations by selecting students at-risk with elevated *T*-scores on the Child Behavior Checklist (Achenbach, 1991) for externalizing and internalizing behavior as well as a comparison group of students. The authors found significant moderation correlations between teacher ratings of externalizing behavior and observations of aggressive and noncompliant behavior (r = .52) while there was no correlation between teacher ratings of internalizing behavior and observations of withdrawal and isolation (r = -.12). Limitations of this study include a small sample of internalizing students (n = 9), completing the observations in an artificial play environment instead of the natural contextual environment of the classroom, and observers may have missed behavior because scan sampling was used to assess six behaviors over a large number of students.

Students themselves are the best informants for rating internalizing symptoms while parents and teachers are ideal for rating externalizing behavior (Smith, 2007). Although, there are noted concerns with relying on teacher informants for student internalizing symptoms, many screening systems initially rely solely on teacher identification (Severson et al., 2007).

Although informant discrepancies are common they may not be error and instead may offer unique and valuable contextual information as a child's behavior is expected to vary in different situations (Dirks et al., 2012). De Los Reyes and colleagues (2015) evaluated studies that examined validity of multi-informant assessment and offered evidence for the incremental and construct validity of multi-informant assessments, which suggested that differences offered important clinical information and a multi-informant approach was warranted. De Los Reyes (2013) developed the Operations Triad Model as a framework for interpreting rater inconsistencies, which assumed that inconsistencies were *diverging* and differed for meaningful reasons or *compensating* where differences were due to measurement error. To determine if inconsistencies are diverging or compensating, De Los Reyes and colleagues (2015) recommend independently observing student behavior or completing more complex assessments and/or interviews to guide diagnosis and treatment.

Using assessment results from one source provides a narrow view of student behavior and could offer untrustworthy results when making decisions based on behavior. Ultimately, many conclude that the differences offer valuable information that should remain in context and it is recommended to gather rating scales from multiple informants to provide trustworthy results and understand student behavior in different contexts (De Los Reyes, 2005; Dirks et al., 2012; Smith, 2007). While this may be a feasible recommendation for completing behavioral

assessments on a small population of high needs students, a multi-informant assessment process would be difficult in the case of universal screenings where all students are briefly rated.

Teacher Differences in Ratings

Several articles examined teacher differences in ratings of student behavior to understand how teachers interpreted rating scales differently through applying generalizability theory, which concurrently assessed multiple facets of variance (Briesch, Swaminathan, Welsch, & Chafouleas, 2014). DBR researchers have examined multiple facets of variance in ratings in great depths. On teacher ratings of prosocial behavior, such as resolves conflicts and interacts cooperatively, 28-41% of the variance in direct behavior ratings for preschool students was credited to rater effects rather than to the child and their behavior with a sample of four teachers rating 15 students (Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese, 2007). However, follow-up generalizability research with two teachers and two observers rating middle school students on behaviors of academic engagement and disruptive behavior found only 2-5% of the variance was due to rater effects but 40-43% of the variance was unexplained (Chafouleas et al., 2010). Briesch, Chafouleas, and Riley-Tillman (2010) also completed a generalizability study with two teacher raters and reported that 29.5% of the variance in behavior ratings was attributed to rater effects with 20% of that variance deriving from rater bias. While the generalizability DBR studies raise concerns regarding differences in teacher ratings, the external validity of the studies is limited due to the low sample size of teachers and students included in the three studies.

Correlations between DBRs and standard direct observations resulted in approximately 45% of the variance in ratings being due to rater bias, with raters underestimating positive behavior and overestimating negative behavior (Christ, Riley-Tillman, Chafouleas, & Jaffery, 2011). The accuracy of behavior ratings varies in relation to low, medium, or high rates of

behavior, which suggests that teacher accuracy in rating may be compromised when completing classwide ratings with variability in the frequency of behaviors displayed (Harrison, Riley-Tillman, & Chafouleas, 2014). However, Harrison et al. (2014) and Christ et al. (2011) both completed studies with undergraduate teachers and brief video tapes of behavior, which makes the generalizability of findings to naturalistic settings unknown. Additionally, all of the DBR studies involved low teacher and student sample sizes and the results may not apply to rating scales because teachers completed brief ratings of behavior in a specified time period rather than rating for the previous month(s). In spite of the limitations, the DBR literature suggests that there are differences in the way teachers rate behavior and the frequency of behavior is relevant to consider.

Other studies have also examined the facet of the rater with more traditional rating scales that rely on longer periods of time (i.e. "consider the last four weeks") for rating scales. For example, in one study, teacher pairs completed the externalizing subscales of the Behavior Assessment System for Children-Second Edition-Teacher Report Form (Reynolds & Kamphaus, 2004) and the Achenbach System of Empirically Based Assessment Teacher Report Form (Achenbach & Rescorla, 2001) and found 12-16% of the variance was attributable to the classroom (Bergenson, Floyd, McCormack, & Farmer, 2008). The conclusions were strengthened by large sample sizes for generalizability studies of 12 teachers and 61 students but limited because a partially nested design was necessary given that all teachers were unable to rate all students (Bergenson et al., 2008). Using a Brief Behavior Rating Scale, Minor (2013) examined variability due to the rater on cooperation and communication items from the Social Skill Improvement Scale (Gresham & Elliott, 2008) assessment in a preschool classroom with two raters rating six students. The authors found that 38-42% of the variance was attributable to

facets related to the rater with higher levels for communication compared to cooperation but the external validity is limited due to the sample of preschool at-risk students, small sample size with two teachers rating six students, and only using portions of a rating scale. Taken together, the generalizability research suggests that the rater should not be ignored in the context of behavior rating scales.

In addition to generalizability studies, the rater variable has been analyzed by examining between-rater variance with multilevel modeling. Mashburn, Hamre, Downer, and Pianta (2006) examined rater effects by calculating the intraclass correlation coefficient teacher ratings for preschool children on the Teacher Child Rating Scale (Hightower et al., 1986), a measure of social competence and behavior problems. Results indicated that between 15-33% of the total variance was explained by the teacher for ratings of preschoolers' behavior. Mashburn et al. (2006) also found that fewer years of experience, higher self-efficacy, non-Caucasian race/ethnicity, and lower child-teacher ratios were related to ratings of child behavior and social competence. However, the study was completed by preschool teachers and it is unknown if differences between teachers are due to actual student behavior or perceptions because teachers did not rate the same students. The research reviewed suggests that teachers rate students differently. Such differences in teacher ratings will impact how students are identified through universal screening and may over or under identify based on the classroom.

Theoretical Frameworks

The aforementioned differences between raters and informants have many theorized causes. The most parsimonious hypothesis is that there is situational specificity that depends on the context for the rating and students may act differently in different settings (Achenbach et al., 1987; Elliott et al., 1993; Kazdin, 1979). De Los Reyes and Kazdin (2005) expand on this

hypothesis and suggest the Attribution Bias Context (ABC) model as a theoretical foundation for understanding rater differences. The ABC model dictates that discrepancies between informants are caused by the context, the informant's attributions of the causes of the behavior (Jones & Nisbitt, 1972), and informant perspectives in terms of goals of the assessment and preferred treatment. To assist in the reduction of discrepancies, De Los Reyes and Kazdin (2005) encourage future research to investigate the impact of directing raters to use the same heuristic and systematic processes for accessing memory information about the child's behavior. Lastly, Wilson and Bullock (1989) posit that raters have different frames of reference, which causes raters to have different perceptions and understandings of the behavior they are rating (Barkley, 1988). Understanding theories behind causes in rater differences can offer solutions to reducing differences and improving accuracy.

Synthesis

Research indicates informant discrepancies are often present with differences attributed to type of behavior and purpose of the assessment. However, the informant discrepancy literature focuses on full length assessments of at-risk children rather than universal screening of all students. Two studies did examine universal teacher rating or nomination of student depressive behavior but are limited by the lack of validity and reliability evidence for the identification methods and the low rates of student consent returned (Auger, 2004; Cunningham & Suldo, 2014). Another noted accuracy concern with teacher ratings is that a sizable portion of the variance in behavior ratings is often due to the rater. Only one of the generalizability studies used a traditional rating scale and found less variance due to the rater (Bergenson et al., 2008), but the study was limited by using a partially nested design for a generalizability study. The evidence from brief behavior rating scale generalizability studies (Chafouleas et al., 2007;

Briesch et al., 2010; Christ et al., 2011; Minor, 2010) suggest that the rater is a large factor in the measurement of child behavior. Such differences impact how students are identified to receive additional social and emotional supports. Differences in ratings may be due to the context that the behavior occurs for each teacher and/or as the ABC model (Jones & Nisbitt, 1972) suggests, teacher perceptions and attributions of the cause of the behavior. There may be benefits to developing a common method of rating to increase objectivity.

Reducing Measurement Error

Training teachers is one way to reduce variability due to the teacher and teacher bias. Training raters should assist then in identifying behaviors, using a system, and making a rating in reference to a standard or criterion (Spool, 1978). Hoyt and Kerns (1999) found that training raters reduces but not eliminates rating concerns. According to the Organizational Consultation literature, training raters improves accuracy and reduces differences in rating style (Sulskey & Balzer, 1988; Stamoulis & Hauenstein, 1993; Woehr & Huffcutt, 1994).

The Systematic Screening for Behavior Disorders (Walker et al., 1988) includes a multiple gating process that relies heavily on teachers noticing symptoms for nominations. Teachers are trained on recognizing internalizing and externalizing symptoms to aid in identification, but the effect of training teachers for this assessment system has yet to be empirically examined (Tomb & Hunter, 2004). Moor and colleagues (2007) examined a teacher training for identifying depressive symptoms in adolescents through a pre-post randomized design of a psychoeducational teacher training on recognizing depressive symptoms, and found a slight decrease in sensitivity for the training group (52% at pretest and 45% at posttest) while the control group remained stable (41% to 43%). The results do not suggest that training could be beneficial, but the screening measure was created by the authors and the validity and test-retest

reliability evidence is unknown and the training provided education on depression rather than how to rate students. Several sources suggest teacher training as best practice prior to completing universal rating scales but offer no empirical evidence that training teachers improves results (Severson et al., 2007; Walker, 2010; Lane, 2012; Weist et al., 2007). Overall, despite the recommendation as a best practice, there is a lack of research on training teachers in mental health screening.

Direct Behavior Rating and Training

The DBR researchers have studied the effect of training teachers to improve accuracy and confidence in teacher ratings for brief and targeted ratings of behavior. Schlientz, Riley-Tillman, Briesch, Walcott, and Chafouleas (2009) examined practice and performance feedback training compared to a brief familiarization training and found that the training condition resulted in greater differential accuracy with undergraduate students rating 3-minute video clips of student behavior with difference scores on academically engaged of 18.86 for the training condition and 24.65 for the brief familiarization condition (d = -0.89). In contrast, LeBel, Kilgus, Briesch, and Chafouleas (2010) compared training intensities for teachers rating video clips and determined that the most intensive and direct training did not improve accuracy and there was no difference on disruptive behavior direct training, indirect training, or no training ($p > .05$), and improved accuracy on academic engagement for indirect training compared to direct training (d = 0.94).

Training improved ratings for disruptive behavior with undergraduate students rating 1-minute videos of student behavior (*partial* η^2 = .006) but not academic engagement or compliance (Harrison et al., 2014). Chafouleas and colleagues (2012) evaluated six different training conditions with undergraduate students rating video clips of student behavior. The authors overall found that Frame of Reference training improved absolute difference accuracy (d

= 0.80 to 1.03) compared to standard familiarization training. Lastly, Chafouleas et al. (2015) completed an initial evaluation of a web-based training system that included modeling with Frame of Reference training and opportunities to practice with immediate correct feedback for undergraduate students rating video clips and found that the training reduced difference scores with expert raters (*partial* η^2 = .38) but results depended on target of behavior and duration of behavior exhibited.

The DBR training studies are limited by using artificial video clips of student behavior rather than behavior from a naturalistic setting where teachers are interacting with and rating all students. Furthermore, only one study (LeBel et al., 2010), used teachers as participants instead of undergraduate teachers and this study found that training did not improve accuracy. These limitations reduce the generalizability of the findings to a normal educational setting with universal screenings of student behavior. The ideal training format and level of intensity has yet to be identified within the DBR literature but taken together, results suggest that at least an indirect training is beneficial and a more intensive training, such as frame of reference, training may be appropriate for more difficult to rate behaviors.

Types of Training

Below I will discuss two common methods for rater training. First, familiarization training will be reviewed and then a more intensive rater training, Frame of Reference training, will be examined. Due to the paucity of research on training raters in the school psychology literature, the organizational consultation research is reviewed to identify ideal training methods for rater training.

Familiarization training. Familiarization training is the minimal level of training required for screening (Sevenson et al., 2007). Familiarization training involves indirect

instruction of material and lacks practice, participation, and corrective feedback (Harrison et al., 2014). The DBR training literature discussed above included familiarization training conditions in some studies. One study compared a familiarization training to an intensive training and found significant results for the intensive training improving the accuracy for disruptive behavior ratings, *partial* $\eta 2 = .006$, but nonsignificant results for academic engagement and compliance behavior ratings (Harrison et al., 2014). The authors suggested that familiarization training may provide acceptable levels of accuracy (Harrison et al., 2014). In an organizational consultation study, Lievens and Sanchez (2007) examined an intensive training condition compared to an indirect training for consultant competency modeling and found that the intensive training produced greater accuracy, *partial* $\eta^2 = .28$.

Frame of Reference. Frame of Reference (FOR) training is frequently cited as the preferred method of training raters in the organizational consultation literature and is often studied in connection to employee performance appraisals (Roch et al., 2012). As a more intensive training method than familiarization training, FOR training assists raters in arriving at a common understanding of performance when rating behavior (Woehr, 1994), and has been shown to improve accuracy by showing samples of behavior that will be rated, offer standards of comparison, and/or provide objective or subjective weights for specific behavior to create a standard reference to use when rating (Sulskey & Balzer, 1988). Moreover, FOR training often includes practice and feedback during training in comparison to standards (Roch et al., 2012). FOR training differs from Rater Error training because there is a lack of emphasis on changing the distribution or scores to prevent rating errors (leniency, severity, halo, central tendency; Roch et al., 2012).

In a meta-analysis of different training methods, Woehr and Huffcutt (1994) examined 29 articles and reported the highest effect size for FOR training ($d = 0.83$) compared to other rating training methods. However, caution should be used when interpreting results because only six studies evaluated FOR training. Roch and colleagues (2012) updated the 1994 study with another meta-analysis of rater training with 23 studies. The researchers found an effect size of $d = 0.50$ sizes on types of accuracy corresponding to expert ratings with a larger effect sizes for differential accuracy ($d = 0.77$) and recall/behavioral accuracy ($d = 0.88$) (Roch et al., 2012).

FOR training has been studied in relation to ratings of social skill and anxiety behavior by studying the effect of differing intensities of FOR training for ratings of adults (Angkaw, Tran, & Haaga, 2006). The authors assessed interrater reliability across training conditions to examine the effect of training and the authors found that FOR training reduced discrepancies social skills, ($p < .001, f = .92$) and anxiety ($p < .001, f = .63$), but there was no significant difference between moderate and intensive FOR training on anxiety. The effect of FOR training on biased personality-based job ratings found moderate effect sizes for both administrative assistants ($d = 0.44$) and supervisors ($d = 0.68$; Aguinis, Mazurkiewicz, & Heggestad, 2009). Neither of the studies mentioned above took place in a school or with rating children so the applicability to universal teacher ratings in school is unknown.

Synthesis

Previous research findings on training teachers on behavior ratings have been inconsistent. Some studies found no effect (Moor et al., 2007; Lebel et al., 2010) while others found improvements with rater training (Schlientz et al., 2012; Chafouleas et al., 2012; Chafouleas et al., 2015) or mixed results (Harrison et al., 2014). However, Moor and colleagues (2007) created a rating scale for their study and did not provide reliability or validity evidence

for the scale, which reduces confidence in the results because it is unknown if the measure provided consistent results and truly measured depressive symptoms. The DBR studies are limited by the lack of external validity to natural school settings with universal ratings of student behavior over several weeks. When examining organizational consultation literature, research consistently found FOR training to be an optimal training to improve rater accuracy (Roch, 2012; Woehr 1994; Angkaw, 2006; McIntyre, Smith, & Hassett, 1984; Uggerslev & Sulskey, 2008). However, these studies involve rating adult behavior in employment settings. Therefore, effect of FOR has yet to be studied in a naturalistic school setting with teachers rating their classroom for the purpose of identifying students for intervention.

Purpose

A review of the behavior rating research raises concerns regarding the accuracy of teacher ratings of student behavior, which is required for universal SEB screening. Inaccurate results and teacher biases can cause misidentification in students identified for SEB supports. Such misidentification can cause schools to misallocate resources in providing interventions. Training teachers prior to completing screening ratings offers a potential solution to mitigate biases and inaccurate ratings. The current study expands on the DBR research by applying examinations of rater training to universal SEB screening of behavior in a naturalistic setting. The present study aims to examine the effect of training teachers on the accuracy and distribution of SEB screening results to help schools make informed decisions about implementing SEB screening. Thus, the study answers the following research questions:

Research Questions:

1. What is the effect of an intensive teacher training on the relationship between teacher ratings on universal screenings and independent systematic direct observations for

academic behavior?

2. What is the effect of an intensive teacher training on the relationship between teacher ratings on universal screenings and independent systematic direct observations for social behavior, including prosocial and disruptive behavior?
3. What is the effect of an intensive teacher training on the distribution, number of students identified, and scores for SAEBRS teacher ratings of Total Behavior?

Given the review of research, it is hypothesized that the intensive teacher training condition will lead to lower differences between the SAEBRS and SDOs of behavior. Additionally, the training is not expected to impact the distribution, number of students identified, or Total Behavior ratings.

CHAPTER III: METHODS

The study collected data with 64 teachers and used a randomized design to address the research questions. See below for information regarding methods specific to answer each research question.

Participants

This study took place in four schools from two school districts in a Midwestern state. One school was a small rural district (District A) and the other was a large suburban district (District B) with three schools included. According to Missouri Department of Elementary Secondary and Education School Report Cards (2017), students in District A were predominately white (94.8%) and approximately half of them (52.4%) were eligible for a free or reduced priced lunch (FRL). Participating schools in District B were more diverse. The first school enrolled 49.6% Black students, 6.2 Hispanic students, and 32.3% White Students (100% FRL). Students in the second school were 19.7% Black, 7.6% Hispanic, and 60.4% White, with just over half (59.4%) being eligible for a FRL. The third school enrolled 17.7% Black students, 6.5% Hispanic students, and 64.6% White students, with 58.1% being eligible for a FRL.

The primary participants were 64 general education teachers for grades kindergarten-8th. Although demographic data were not collected for individual teachers, teachers were predominately Caucasian and gender was estimated based on name as 98% female. The teachers rated 1158 students during universal screening, 16% of whom were in kindergarten, 16% in first grade, 18% in second grade, 16% in third grade, 17% in fourth grade, 16% in fifth grade, and .17% in eighth grade.

A total of 74 students were selected from the 1158 who were rated to be observed. Only one observed student was in the At Risk range for District A, while 27 students (42%) in District

B were rated in the At Risk range. Therefore, Total Behavior scores were higher for District A compared to District B. The majority of observed students were in grades kindergarten through fifth grade, while two eighth grade students were observed in District A. Individual demographic data were not collected for the observed students but gender was categorized based on name and over half of the sample was classified as male (58.1%).

Measures

The criterion variables for the study were difference scores between the Academic and Social Behavior scales of the Social, Academic, and Emotional Behavior Risk Screening Scale and systematic direct observations of academic engagement, prosocial behavior, and disruptive behavior. Both measures are described below.

Social, Academic, and Emotional Behavior Risk Screening Scale (SAEBRS)

The Social, Academic, and Emotional Behavior Risk Screener (SAEBRS; Kilgus & von der Embse, 2014) was selected as the screening measure by the school district to identify students in need of social and emotional supports. The SAEBRS is a universal teacher rating scale that consists of 19 items with both strength and deficit worded items, where teachers rate using a four-point behavior frequency scale (0 = never, 3 = always). When rating items, teachers are asked to consider the students' behavior over the past 4 weeks (Kilgus et al., 2016). The scale provides a Total Behavior Score as well as subscale scores for Social Behavior, Academic Behavior, and Emotional Behavior. Across each of these scales, students may be classified as either At-Risk or Not At-Risk (Kilgus et al., 2016). The Total Behavior score has demonstrated strong internal consistency (α = .93-.94) while the subscales were in the adequate range (α = .83-.94; Kilgus et al., 2016). Concurrent validity coefficients ranged from r = .61-.93. Teachers can complete the scale in 1-3 minutes per student (Kilgus et al., 2016). Data consisted of summed

scores with risk cut-offs for At Risk and Not At Risk. Summed scores of 36 or less for Total Behavior, 9 or less of Academic Behavior, 12 or less for Social Behavior, and 16 or less for Emotional Behavior indicate At Risk status (Kilgus et al., 2016).

Systematic Direct Observation (SDO)

Systematic direct observation (SDO) offers an objective method to assess student behavior. Assessment occurs via measurement of specific behaviors that are operationally defined a priori and the use of standardized coding procedures (Hintze et al., 2008). SDO protocols were created with momentary time sampling using a 15-second interval over 15 minutes. Momentary time sampling determines a time period, divides the observation into equal intervals, and assesses if the behavior occurred at the specified interval to provide an overall percentage of behavior (Merrell, 1999).

In this study, every 15-seconds, the observer separately recorded if the student was academically engaged, demonstrating prosocial behavior, and demonstrating disruptive behavior. The SDO form, see Appendix A, provided operational definitions of each behavior with examples. Academic engagement is actively or passively participating in the classroom activity, such as writing, raising hand, answering a question, talking about a lesson or task demand, listening to the teacher, reading silently, or looking at instructional materials. Prosocial behavior is interacting with peer(s) in a positive or neutral way such as working cooperatively, joining task sharing, and/or talking/listening to peers related to classroom instruction and during unstructured nonacademic time, the student is polite and engaging with peers or teacher on appropriate topics and making appropriate responses such as not putting down others, turn taking, eye contact, and nonverbals. Disruptive behavior is action that interrupts regular school or classroom activity or inappropriate interactions such as being out of seat, playing with objects,

acting aggressively, talking/yelling about things that are unrelated to classroom instruction, and/or defiance/noncompliance. Observers recorded if there was no opportunity for academic engagement (e.g., during transitions) or prosocial behavior (e.g., during teacher lecture with the expectation to sit quietly). In contrast, an opportunity to engage in disruptive behavior was always presumed.

Training Conditions

Each participating teacher was randomly assigned to one of two training conditions. The conditions are explained below.

Familiarization training

All teachers received a brief training that includes a brief overview of the measure and directions on how to complete the measure. Teachers in the Familiarization Condition only received the brief overview portion of the training, which lasted approximately 15 minutes. The training occurred in a large group format with a PowerPoint presentation (see Appendix B) reviewing the purpose of the measure, how to complete the forms, and encouraging the consideration of student behavior over the past 4 weeks.

Frame of Reference Training with Familiarization Training

The FOR training followed the Familiarization training for randomly selected participants. See Appendix C for training slides for the FOR intensive training. The FOR training occurred in a large group setting and teachers were grouped by grade level, if possible, to facilitate discussion. Teachers received the same familiarization PowerPoint and then the FOR training portion began. The FOR training focused on creating a common understanding of behaviors and the scaling items. Select items on the scale were reviewed with examples of specific behaviors provided. For example, with Social Behavior, teachers discussed definitions of

disruptive behavior and impulsiveness and brainstormed example behaviors. General guidelines were provided for each scaling anchor, where *never* corresponds to less than once a week, *sometimes* is one to two times per week, *often* is three to five times per week, and *almost always* is several times a day. However, teachers were also encouraged to consider the specific behavior and developmental level of the student when rating the frequency. For instance, when rating distractedness, teachers should consider that the majority of children are distracted for brief periods of time daily so general guidelines may not apply and teachers can instead consider the amount of distracted time or frequency in one day. Similarly, a typical kindergarten student is more likely to engage in distracted behavior compared to a 5th grade student.

Teachers next individually rated an "average" student for their grade level and discussed the ratings and reviewed discrepancies. Researchers provided feedback regarding average student behavior, while also indicating that the average student usually engages in problem behavior *sometimes* and positive behaviors *often*. It was also noted, however, that average is a range of ratings and most students would be characterized as average. The training stressed that extreme ratings (i.e., never or almost always) should be reserved for extreme behavior students, both below and above average. Finally, a vignette described a students' behavior over the past 4 weeks. Teachers completed the rating form for this hypothetical child as a group and reached consensus regarding ratings and explained their reason (frame of reference) to one another for ratings. The trainer provided feedback for the ratings regarding how the teachers interpret behaviors and frequency. At the conclusion of the training, the trainer stressed participants to remember how items were defined, objectively consider how often the behavior occurs, only consider the past month and what is average for their grade, and reserve extreme ratings for

nontypical students. The additional frame of reference training was 30 minutes in addition to the familiarization training.

Procedures

First, Institutional Review Board approval was received and schools were recruited. Figure 1 displays a flow chart of the study procedures for trainings, ratings, and observations. All teachers received the brief familiarization control training at an afterschool meeting. Then, the research project was described and the informed consent was reviewed and teachers had the opportunity to ask questions and then teachers signed consent. Participating teachers completed an observation availability form to identify times when students are typically engaged in academic tasks with the opportunity for peer interaction, such as peer tutoring, academic centers, or group projects. Once the consent process was completed, half of the participating teachers were randomly selected to stay for the additional FOR training through picking names out of a container. The additional training took no longer than 30 minutes.

For district A, teachers completed the SAEBRS on their entire classroom the week after the training while at district B teachers completed the SAEBRS on the day of the training. Teachers rated their entire classroom on the SAEBRS (via electronic survey software) unless the parental opt-out form was returned. After the SAEBRS was completed, the data were reviewed by the school counselor and four students were selected from each participating teacher's classroom, with the goal of receiving at least two returned signed consents per classroom. Students selected were those with the closest scores to the At Risk cutoff for Total Behavior, with two student classified as At Risk and the other two Not At Risk but near the cutoff of 36. The goal was to select students whose likelihood to be selected for interventions may change based on the accuracy of teacher ratings. Parental consents for observations were sent home and

students with returned consents were included in the study, after student assent was obtained. In the event that parental consents were not returned for a classroom, additional students were selected for observations with a larger range of Total Behavior score. Several classrooms did not receive returned consents for two students, even after repeated reminders. Teachers and school counselors reported some confusion and concern with the parental consent form.

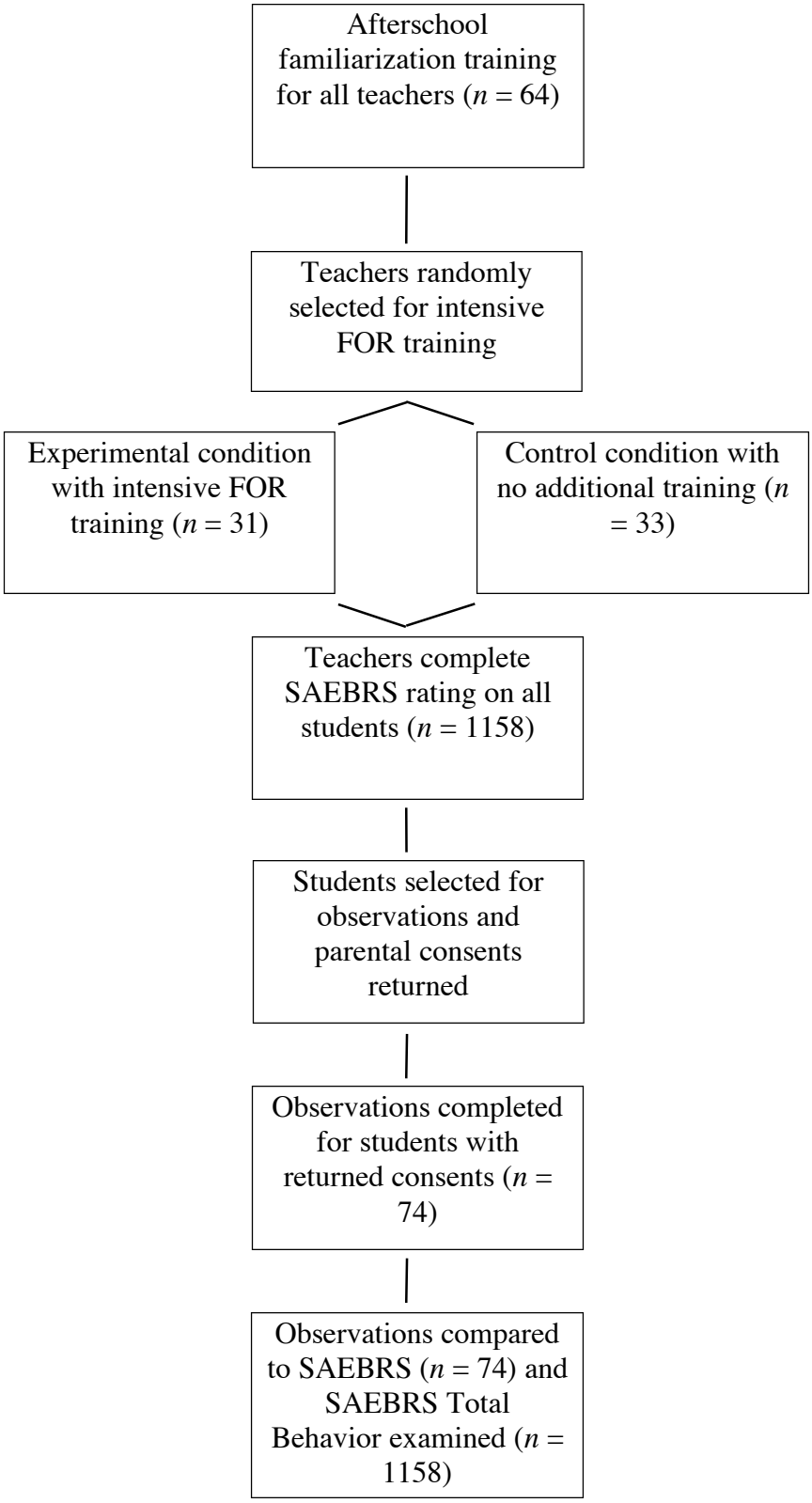


Figure 1. Study Procedures.

Participating students from each classroom were observed one to four times during unstructured instructional time with opportunities to interact with peers. Students from each classroom were observed at the same time with the observer rating the first student and then the second student. Observations occurred after the training and SAEBRS rating. Observations occurred 1 to 16 weeks after the SAEBRS was completed, based on when parental consent was returned and teacher availability. The majority of observations (85%) occurred up to 8 weeks following teachers completing the SAEBRS for their classrooms, while 11% occurred 12 weeks after, and 4% 16 weeks after. At the conclusion of each observation, the percentage of time each student engaged in on-task, prosocial, and disruptive behavior was calculated and the median score for each behavior per student was selected.

Treatment Fidelity

A total of 100% of teacher trainings were observed for fidelity by a school psychology graduate student with a Treatment Fidelity Checklist (Appendix D). The total number of items observed was divided by the total number of items and multiplied by 100. The results of the observation were that all components were included in the observed trainings, which resulted in 100% correct implementation of the trainings.

Observer Training and Interobserver Agreement

Observers were trained through direct teaching of the observation code and momentary time sampling. The training included modeling, practice, and corrective feedback. The primary investigator first reviewed the purpose of the training and expectations for observers (e.g. arrive on time, sign-in at the office). Next, observers learned how to complete momentary time sampling and reviewed the observation form and behavioral operational definitions. The training reviewed when to rate No Opportunity for academic engagement and prosocial behavior. The

group discussed examples of each behavior and reviewed potential classroom activities. Then, the primary investigator modeled completing the observation form for several scenarios and the group practiced watching videos of classrooms while completing the observation form. Observers practiced rating two students at once and had opportunities to ask clarifying questions. Lastly, observers independently completed the observation form for videos of classrooms during academic activities with opportunities for peer interaction. Results were reviewed and corrective feedback was provided. All observers reached reliability of 85% with each other and the primary investigator or co-advisor before beginning observations.

Interobserver agreement (IOA) was calculated for observations by having a second person observe the same student and independently complete the Observation Form. Coefficient kappa values were calculated to determine IOA where the proportion of agreements was subtracted by the proportion of expected agreements and divided by one minus the proportion of expected agreements (Hintze, 2005). IOA was completed for 9% of total observations and 24% of students. Results were $k = .97$ for disruptive behavior (98% agreement), $k = .73$ for academic engagement (96% agreement), and $k = .53$ for prosocial behavior (93% agreement), which corresponds to near perfect agreement for disruptive, substantial for academic engagement, and moderate for prosocial (Hintze, 2005; Landis & Koch, 1977).

Data Analyses

This study incorporated a randomized experimental design. The first two research questions were analyzed with multilevel modeling to compare the SAEBRS teacher ratings to objective SDOs of academic engagement, prosocial behavior, and disruptive behavior and nonparametric analyses were used if multilevel was not warranted based on between-teacher difference. To create the dependent variable, a difference score for SAEBRS and SDOs was

created as a measure of difference in accuracy (Roch et al., 2012; Chafouleas et al., 2012). First, the median observation percentage for each behavior was selected. For prosocial behavior, observations with less than 30 intervals were excluded due to a lack of opportunity to display social behavior. Disruptive behavior percentage was subtracted by 100 to create a percentage of time not engaged in disruptive behavior. Then, observation percentages for each SDO behavior and SAEBRS scores were standardized by school to create z-scores. Difference scores were created for each student by subtracting the median SDO z-score from the SAEBRS z-score and taking the absolute value. The procedure was completed for Academic Engagement SDO and SAEBRS Academic Behavior, Prosocial Behavior SDO and SAEBRS Social Behavior, and Disruptive Behavior and SAEBRS Social Behavior.

A hierarchical estimation was used to fit each dependent variable with the treatment at the level 2 model, where random assignment occurred. Analyses were completed with SPSS Statistics 24 in consultation with multilevel SPSS guides (Peugh & Enders, 2005; Albright, Jeremy, & Marinova, 2010) and full maximum likelihood estimation was used for all analyses. First, the unconditional model was tested for each dependent variable to examine between teacher effects on difference scores.

Level-1 Model (1)

$$Difference\ Academic\ Behavior_{ij} = \beta_{0j} + e_{ij}$$

Level-2 Model

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

Level-1 Model (2)

$$Difference\ Prosocial\ Behavior_{ij} = \beta_{0j} + e_{ij}$$

Level-2 Model

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

Level-1 Model (3)

$$Difference\ Disruptive\ Behavior_{ij} = \beta_{0j} + e_{ij}$$

Level-2 Model

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

Next, intervention status was dummy coded and included as a predictor level variable at Level 2 to determine if intervention significantly predicts the absolute value of the difference score. It is hypothesized that intervention will reduce the grand mean, γ_{00}, and lead to a *p*-value of less than .05.

Level-1 Model (4)

$$Difference\ Academic\ Behavior_{ij} = \beta_{0j} + e_{ij}$$

Level 2 model

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\text{Intervention}_j + u_{0j}$$

Level-1 Model (5)

$$Difference\ Prosocial\ Behavior_{ij} = \beta_{0j} + e_{ij}$$

Level 2 model

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\text{Intervention}_j + u_{0j}$$

Level-1 Model (6)

$$Difference\ Disruptive\ Behavior_{ij} = \beta_{0j} + e_{ij}$$

Level-2 model

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\text{Intervention}_j + u_{0j}$$

Effect sizes are calculated with Hedge's *g* (1982) and interpreted with Cohen's (1988) guidelines where .2 = small, .5 = medium, and .8 = large. If a multilevel approach was unnecessary, as

based on a significant value greater than .05 of the between teacher variance (Peugh & Enders, 2005) and an intraclass correlation coefficient .10 or less (Lee, 2000), then nonparametric Mann Whitney U tests could be completed to compare mean ranks. Tests produce mean ranks for each condition and a *U* value, and significance would be evaluated with a criterion of $p < .05$.

To answer research question three, SAEBRS data for all students in each district were accessed to examine the impact of the intensive training on SAEBRS Total Behavior distribution and ratings. It is expected that the training will not impact the Total Behavior scores or distribution because rater error training (McIntyre et al., 1984) was not included in the teacher training intervention. Homogeneity of variance for both conditions was compared with Levene's Test of Equality of Variances, which yields a *F* statistic. A significant *p*-value of less than .05 for Levene's Test *F* statistic indicates a difference in variance between the two training conditions. A nonsignificant *p*-value is expected because differences in variance would be an unintended consequence of the intervention. Percentage of students at-risk was aggregated by teacher and an independent sample *t*-test compared the classroom percentage of students at-risk by training condition. Again, one expects that the intensive training would not alter the number of students identified as at-risk. Therefore, a *p*-value greater than .05 is expected. Multilevel modeling was used to examine the impact of the intervention for Total Behavior scores while accounting for between rater variance. The intervention is not expected to be a significant (*p*-value greater than .05) predictor variable, reduce the intraclass correlation coefficient, or change the grand mean, γ_{00}. First, the unconditional model was tested to produce an intraclass correlation coefficient of between-teacher variance. In the equation, the Total Behavior Score of student *i* nested in teacher *j*. Then, intervention was added as a predictor variable at level two as an additional fixed effect estimate, γ_{01}, which quantifies the effect of intervention.

Unconditional Level-1 Model (7)

$$Total\ Behavior\ Score_{ij} = \beta_{0j} + e_{ij}$$

Level-2 Model

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

Level-1 Model (8)

$$Total\ Behavior\ Score_{ij} = \beta_{0j} + e_{ij}$$

Level-2 model

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\text{Intervention}_j + u_{0j}$$

CHAPTER IV: RESULTS

The purpose of this study was to examine the effect of the Frame of Reference teacher training on the relationship between teacher ratings on universal screenings and independent systematic direct observations of academic and social behavior. The results are presented based on research question and split into two sections based on type of analyses. First, descriptive statistics are reviewed. Then for research question one, the effect of an intensive teacher training is examined for SAEBRS Academic Behavior compared to academic engagement observations. Next, results are presented for research question two on the effect of an intensive teacher training on the relationship between SAEBRS Social Behavior and prosocial behavior and also disruptive behavior observations. Lastly, results for research question three are reviewed, which examines the impact of an intensive teacher training on the distribution, number of students identified, and ratings for SAEBRS ratings of Total Behavior for all students screened. It is hypothesized that results will not indicate a statistical difference on the distribution, number of students identified, or Total Behavior scores.

Descriptive Statistics

For Research Questions 1 and 2, the sample contained 74 student participants for observations. Table 1 presents grade-level and SAEBRS Total Behavior data for students screened and observed. Table 2 displays grade-level and SAEBRS Total Behavior data for all students rated by teachers for research question three. SAEBRS Total Behavior scores were higher for District A compared to District B, but included a smaller sample size.

Table 1

Student Grade Levels and Total Behavior SAEBRS Scores for Observation Participants

	n	SAEBRS *M* (*SD*)
Grade		
Kindergarten	13	38.4 (7.8)
First	14	38.9 (7.2)
Second	13	37.9 (6.7)
Third	13	37.9 (8.6)
Fourth	10	40.2 (6.7)
Fifth	9	39.9 (10.1)
Sixth	--	--
Seventh	--	--
Eighth	2	57 (0)
Total	74	

Note. Total Behavior At Risk is less than 37.

Table 2

Student Grade Levels and Total Behavior SAEBRS Scores for Screening Participants

	n	SAEBRS *M* (*SD*)
Grade		
Kindergarten	189	45.0 (10.3)
First	183	40.0 (11.8)
Second	207	44.3 (10.3)
Third	187	44.1 (11.3)
Fourth	195	43.2 (11.4)
Fifth	182	41.4 (12.6)
Sixth	--	--
Seventh	--	--
Eighth	15	54.7 (5.2)
Total	1158	

Note. Total Behavior At Risk is less than 37.

Comparisons to Observations of Behavior

The first two research questions inquired about the effect of an intensive teacher training on the relationship between teacher ratings on universal screenings and independent systematic direct observations of behavior. The data to address this question are below and are listed by academic behavior, prosocial behavior, and disruptive behavior.

Academic Behavior

Table 3 displays mean difference scores for Academic Behavior and mean academic engagement percentages. Difference scores were lower for the FOR treatment condition.

Table 3

Observation Descriptive Statistics for the Frame of Reference (FOR) and Familiarization (Control) Conditions

	Academic Behavior			Prosocial Behavior			Disruptive Behavior		
	n	*M* (*SD*)	Mean Difference (*SE*)	*n*	*M* (*SD*)	Mean Difference (*SE*)	*n*	*M* (*SD*)	Mean Difference (*SE*)
FOR Intervention	37	84.83 (16.77)	0.94 (.12)	31	37.98 (32.10)	1.33 (.15)	37	7.24 (13.61)	0.95 (.13)
Control	37	86.44 (16.22)	1.20 (.15)	30	30.57 (26.05)	1.03 (.14)	37	4.27 (8.20)	0.88 (.14)

First, a multilevel model was tested. Results of the unconditional model suggest that a significant amount of the variance in Academic Behavior Difference scores was explained within raters ($Z = 3.4$ $p < .001$), but the intercept parameter indicated that the intercepts did not vary significantly across raters ($Z = .65$, $p = .52$). The intraclass correlation coefficient (ICC) was .153, indicating that 15.3% of the variance in Academic Behavior Difference scores is attributable to differences between raters. The ICC greater than 10% suggests that there is a substantial portion of the variance due to between-teacher differences (Lee, 2000). Therefore, multilevel modeling is warranted despite the nonsignificant between-rater *p*-value. When intervention was added as a level-2 predictor, the effect of intervention was not significant $t(46.69) = -1.53$, $p = .132$. The grand mean for intervention was -.31, with a standard error of .20 ($sd = .54$). FOR intervention classrooms had slightly lower Academic Behavior Difference scores, although the difference was not significant in a multilevel model. Effect size was calculated and found a small effect (Hedge's $g = .31$) for the intensive frame of reference condition ($m = .94$, $sd = .75$) compared to the control condition ($m = 1.20$, $sd = .90$).

Prosocial Behavior

Table 3 displays mean difference scores for Prosocial Behavior and mean percentage of observed prosocial behavior. Difference scores were lower for the Familiarization Control condition and included a smaller sample size due to excluding observations with less than 30 intervals of prosocial behavior opportunities. First, a multilevel model was tested, results of the unconditional model suggest that a significant amount of the variance in Prosocial Behavior Difference scores was explained within raters ($Z = 3.8$ $p < .001$), but the intercept parameter indicated that the intercepts did not vary significantly across raters ($Z = .43$, $p = .69$). The ICC was .079, indicating that 7.9% of the variance in Prosocial Behavior Difference scores is

attributable to differences between raters. It was determined that multilevel analyses were unnecessary due to the nonsignificant between-teacher *p*-value and ICC less than 10%, which indicates trivial Level 2 effects (Lee, 2000). Instead, normality was tested with Shapiro Wilk's *W* test to determine if parametric analyses are appropriate and found that the normality assumption was violated for intervention, Familiarization Control $W(30) = .94, p = .07$, FOR Intervention $W(31) = .92, p = .02$. An independent-samples Mann Whitney U test was completed to compare mean ranks for training conditions. Results found no significant difference between training conditions for Prosocial Behavior, $U = 560, p = .17$, with a mean rank of 27.8 for the Familiarization Control condition and 34.1 for the FOR intervention condition. Effect size was calculated and found a small effect (Hedge's $g = -.36$) for the control condition ($m = 1.03$, $sd = .79$) compared to the intensive frame of reference condition ($m = 1.33$, $sd = .86$).

Disruptive Behavior

Table 3 displays mean difference scores for Disruptive Behavior and mean percentages of observed disruptive behavior. Difference scores were lower for the Familiarization control condition. First, a multilevel model was tested, results of the unconditional model suggest that a significant amount of the variance in Disruptive Behavior Difference scores was explained within raters ($Z = 4.2$ $p < .001$), but the intercept parameter indicated that the intercepts did not vary significantly across raters ($Z = .64, p = .52$). The ICC was .108, indicating that 10.8% of the variance in Disruptive Behavior Difference scores is attributable to differences between raters. The ICC greater than 10% (Lee, 2000) suggests that there is a substantial portion of the variance due to between-teacher differences. Therefore, multilevel modeling is warranted despite the nonsignificant between-rater *p*-value. When intervention was added as a level-2 predictor, the effect of intervention was not significant, $t(53.94) = .316, p = .75$. The grand mean for

intervention was .06, with a standard error of .20 (*sd* = .53). FOR intervention classrooms had slightly larger Disruptive Behavior Difference scores in a multilevel model, although the difference was not significant, indicating no difference on comparisons of SAEBRS Social Behavior and SDOs of Disruptive Behavior. Effect size was calculated and found a negligible effect (Hedge's *g* = -.08) for the control condition (*m* = .88, s*d* = .85) compared to the intensive frame of reference condition (*m* = .95, *sd* = .82),

Impacts on SAEBRS Total Behavior Ratings

The third research question inquired about the effect of an intensive teacher training on the SAEBRS teacher ratings of Total Behavior including number of students identified as At Risk, homogeneity of variance, and Total Behavior scores, although significant differences were not expected based on the focus of the intervention. Table 4 presents descriptive statistics for SAEBRS Total Behavior by condition. Means and standard deviations are similar but the range and number of students identified At Risk is lower for the FOR treatment condition. Results of an independent samples *t*-test indicate a non-significant difference $t\ (62) = .554$, $p = .582$, (Cohen's $d = .14$) on percentage of students identified as At Risk by classroom. Levene's Test for Equality of Variance suggests no impact, $F\ (1, 1156) = 3.062$, $p = .080$, from the FOR intervention on the distribution of Total Behavior scores. Therefore, the FOR Intervention did not significantly impact distribution or number of students identified.

Table 4

Total Behavior Descriptive Statistics by Condition

	n	*M*	*SD*	Range (Min-Max)	*M* % of Students as At-Risk
FOR Intervention	567	42.99	11.13	43 (14-57)	24.22%
Familiarization Control	591	43.20	11.79	48 (9-57)	26.56%

Note. Total Behavior At Risk is less than 37.

A multilevel model was used to compare treatment effects for SAEBRS Total Behavior. The unconditional model suggests that a significant amount of the variance in Total Behavior Scores was explained within raters ($Z = 23.4$ $p < .001$) and the intercept parameter signified that the intercepts varied significantly across raters ($Z = 4.3$, $p < .001$). The ICC was .166, indicating that 16.6% of the variance in Total Behavior scores is attributable to differences between raters. When intervention was added as a level-2 predictor, the effect of intervention was not significant t (62.64) = .080, p = .94. The grand mean for intervention was .11, with a standard error of 1.32. FOR intervention classrooms had slightly higher Total Behavior Scores in a multilevel model, although the difference was not significant.

CHAPTER V – DISCUSSION

The previous section presented results for the research questions. The final chapter will review and discuss the results and contextualize results within previous literature and limitations of the study.

Summary of Findings

The current study sought to determine the impact of an intensive teacher training on the accuracy of teacher ratings of SEB functioning. Objective SDOs of behavior in the classroom served as an indicator of "true" student behavior and SAEBRS-SDO difference accuracy scores were derived. As discussed in the literature review, Frame of Reference training was the preferred method of rater training (Roch et al., 2012), and was the treatment used in this study.

Difference from Observation

The first research question examined training impact on Academic Behavior SAEBRS ratings when compared to SDOs of academic engagement. Academic Behavior on the SAEBRS measures a student's behavior in relation to academic activities and includes specific behaviors of academic engagement, preparedness, work production, distraction, and interest in academic material (Kilgus, Eklund, von der Embse, Taylor, & Sims, 2016). Mean academic engagement SDOs median percentages were similar for each condition but when compared to teacher ratings on SAEBRS Academic Behavior, the FOR intervention condition difference accuracy score was lower (thus indicating greater accuracy). Although academic behavior differences were reduced for the intervention condition, the grouping variable was not significant in the multilevel model.

The second research question assessed the effect of the FOR training on accuracy of SAEBRS Social Behavior ratings when compared to SDOs of prosocial behavior and disruptive behavior. Social Behavior on the SAEBRS measures behaviors that relate to maintaining age

appropriate relationships with peers and adults. Teachers rate both positive behaviors (e.g., cooperation with peers, polite and socially appropriate responses) as well as problem behaviors (e.g., arguing, disruptive behavior). Prosocial behavior SDOs served as an indicator for the positive behavior measured on the SAEBRS. Prosocial behavior average percentages were lower for the control condition, however, prosocial behavior included a smaller number of observations due to excluding observations with less than 30 intervals (7 minutes) of opportunities for prosocial behavior. Observations occurred during academic activities, which limited the opportunities for prosocial behavior. Thus, the number of observations was decreased and there were fewer observations to create a median score. The mean difference score was lower for the familiarization control condition, although reductions were not significant on the nonparametric analysis. Results do not suggest a positive or negative impact from the FOR training for prosocial behavior.

While prosocial SDO measured positive social behavior, disruptive behavior SDO measured problem behavior in the classroom. In general, percentages for disruptive behavior were low, with many students engaged in disruptive behavior less than 5% of the observation period. Students within the control familiarization condition displayed disruptive less than those students in the FOR condition. When comparing SDO scores compared to SAEBRS Social Behavior scores, difference accuracy scores were lower for the familiarization control condition, but intervention was not significant in the multilevel model. Thus, impacts from FOR teacher training on rating problem behavior are not conclusive.

The analyses for research questions one and two followed the recommended method of assessing rater training improvements by comparing ratings to objective expert ratings/observations or true scores (Murphy & Balzer, 1989; Roch et al., 2012). The results

suggest no significant improvement in screener rating accuracy when compared to observations, which is surprising given the large body of research supporting FOR training. A meta-analysis of FOR rater training (Roch et al., 2012) found that 56 out of 57 identified studies compared ratings to true scores and found overall moderate to large effects for FOR training. Similarly, an education-based DBR study found the greatest support for FOR teacher training with large effects for absolute difference accuracy (Chafouleas et al., 2012). Harrison and colleagues (2014) noted improvements after practice and feedback training for disruptive behavior but not academic engagement or compliance behavior. However, the DBR studies were not conducted in a naturalistic setting with teachers rating their classrooms over a month for early identification purposes.

The current study did not find improvements on difference accuracy, although the design may have limited the ability to detect differences in accuracy. Examining the accuracy of a teacher rated universal screening measure is difficult because teachers rate their entire classroom and consider behavior over the past month. There are few accepted and feasible measures of a universal screening "true score" as recommended by the rater training literature (Roch et al., 2012). Lower difference scores for academic engagement were found for the intervention condition with a small effect, although differences were not significant. Difference scores for prosocial and disruptive behavior were lower for the familiarization control condition but differences were small and not significant. The lack of improvement in the hypothesized direction for Social Behavior may be due to the lack of opportunity to observe prosocial behavior. Further, the SAEBRS Social Behavior rating was likely influenced by student negative behavior more than prosocial behavior as the positive Social Behavior items do not load highly on the Social Behavior subscale (von der Embse, Pendergast, Kilgus, & Eklund, 2016).

However, it is difficult to detect disruptive behavior with momentary time sampling because the behavior occurs infrequently (Hintze et al., 2008) and the low rates in the current study are likely an underestimation of actual disruptive behavior. Results should be viewed as preliminary in light of the limitations discussed below.

Impacts on SAEBRS Total Behavior Ratings

The study also sought to determine the impact of an intensive teacher training on the distribution of teacher Total Behavior ratings on a universal screening measure. It was hypothesized that there would be no impact on the distribution, number of students identified At Risk, and Total Behavior Scores. The familiarization group and FOR intervention conditions were similar when compared on Total Behavior ratings. There was no significant difference in the percentage of students identified, Levene's test indicated homogeneity of variance, and intervention was not significant in the multilevel model and did not account for a noticeable difference in variance. Although there was not a significant impact, the training condition's range and standard deviation were smaller for Total Behavior. The intervention did not significantly alter Total Behavior scores or distributions; yet, it should be noted that the main goal was to improve accuracy instead of change the distribution of scores. Changes in distribution and Total Behavior scores would be an unintended consequence because the training focused on increasing objectivity rather than altering the distribution of scores.

Rater characteristics are associated with variation in ratings, such as leniency, severity, and central tendency errors (Lumley & McNamara, 1995). Rater training, especially rater error training, is suggested as a means to reduce variation in ratings (McIntyre et al., 1984), although the method has since fallen out of favor. The meta-analysis by Roch et al. (2012) found only four studies published after Woehr and Huffcutt's (1994) meta-analysis that examined distributional

differences such as halo, leniency, and raw means whereas the majority of studies compared ratings to true scores or expert ratings. Even though the method is not widely used, previous research found that rater training impacts rating variation (McIntyre et al., 1984) and is associated with reduced mean ratings and standard deviations (Schlientz et al., 2009). Results of the current study did not find such differences in distribution and variation.

It should be noted that differences in distribution are a potentially flawed rater training outcome measure because differences can also be due to actual classroom differences in ratee behavior and environment (e.g. classroom climate, teaching style, and behavior of other students; Roch et al. 2012; Thomas, Bierman, & Power, 2011; Werthamer-Larsson, Kellam, & Wheeler, 1991). Instead, it is accepted that rater training studies should compare ratings to objective measures of behavior such as expert ratings/observations or true scores (Murphy & Balzer, 1989; Roch et al., 2012; McIntyre et al., 1984). Therefore, research questions 1 and 2 are a better indicator of training impacts, while research question 3 considers potential unintended outcomes of changing the distribution of scores.

Implications for Practice

The current study highlights the importance of considering teacher accuracy of SEB ratings. Although results did not suggest statistically significant improvement for an intensive training, schools should continue to follow screening best practices (Severson et al., 2007; Walker, 2010; Lane, 2012; Weist et al., 2007) and deliver a brief familiarization training prior to universal screening. Familiarization training includes an overview of the process and how to complete the measure but does not involve modeling, practice, or feedback (Harrison et al, 2014). In the current study, a large portion of the variance in multilevel modeling for Total Behavior was due to differences between teachers. Schools should be mindful of between-

teacher differences and complete screening follow-ups to determine if differences are due to classroom differences in behavior or teacher perception (De Los Reyes et al., 2015). Screening follow-ups can also help school personnel decide if intervention is warranted and determine type of intervention (Merrell, 2001; Severson et al., 2007). Increasing teacher accuracy improves identification of students with SEB needs, which can prevent later mental health difficulties (O'Connell et al., 2009). Schools can take additional steps to improve rater accuracy (e.g. supplementary training, corrective feedback for rating errors) but efforts should be locally evaluated to ensure that additional endeavors are effective and worthwhile of teacher time.

Implications for Theory

According to the ABC model, discrepancies between informants are caused by the context, the informant's attributions of the causes of the behavior (Jones & Nisbitt, 1972), and informant perspectives in terms of goals of the assessment and preferred treatment. De Los Reyes and Kazdin (2005) identified the need to examine the impact of directing raters to use the same heuristic and systematic processes for accessing memory information about the child's behavior. The intensive frame of reference intervention training of the current study aligns with the ABC model that raters are influenced by subjective factors and focused on increasing objectivity and reaching consensus on the rating process. However, the lack of statistically significant differences between the intensive training and control conditions for ratings of academic and social behavior in comparison to objective observations of behavior suggest that objectivity was not significantly improved.

Although the training did not improve accuracy or change score distribution, intraclass correlation coefficients for Total Behavior of all students screened suggests that a large portion of the variance is due to between-teacher differences. The variance suggests that more research is

warranted on contributing factors to rater differences in universal screening. Further research on the ABC model and frame of reference training is needed to conclude that context, informant attributions of causes of behavior, and perspectives are influencing screening results and can be altered with teacher training.

Limitations and Future Directions

While this study expands previous literature, there were several limitations. The design assumes that observations serve as an objective and accurate indicator of true behavior and function as an expert rating. Although observations can serve as an expert indicator of behavior, SDO should follow best practice standards, which include ample opportunities to observe the behavior over a number of observation periods and the use of concrete, specific operational definitions (Hintze et al., 2008). Thus, the current study results should be viewed as preliminary as there are several limitations with the ability of observations to serve as an indicator of accuracy for SAEBRS ratings. First, due to scheduling difficulties, observations took place anywhere from 1 to 16 weeks after SAEBRS completion, whereas the SAEBRS instructs teachers to consider student behavior frequency from the previous month. The observations occurred after the screening due to logistical reasons and a few observations occurred months after the rating due to a delay in receiving parental consents. The delay in observations was necessary but reduced the likelihood that observations are accurate representations of behavior at the time of the SAEBRS ratings. Student behavior may change over time due to several reasons (e.g. time of year, academic material, changes in home environment, SEB intervention) and future research should design a study where observations occur the month before the SAEBRS ratings.

Second, the observations served as a "true" indicator of student behavior over one month. However, students were only observed for a total of 15-60 minutes in select class activities, which does not capture the full range of behavior that is rated on the SAEBRS. A total of six students were observed once and 21 students observed twice. In the future, research should design studies to capture behavior throughout the day over 3-4 weeks.

Observations of social behavior was also a limitation. For instance, there were not often opportunities to observe prosocial behavior, thus, many observations are missing prosocial behavior or there was only one available data point instead of taking the median of three observations (one prosocial observation for 27 students, and 13 students did not have any opportunities for prosocial behavior). Thus, prosocial behavior observations are less reliable and representative of true Social Behavior. Further, the observations contained low base rates of disruptive behavior. Hintze and colleagues (2008) recommend observing low base rate behavior with frequency counts or partial-interval time sampling instead of momentary time sampling. Future studies should design observations in the classroom and during unstructured times such as recess to fully observe prosocial behavior and use alternative methods to assess disruptive behavior.

There were also limitations with the students selected for observations. Only one or two students per classroom were observed to make the study feasible. However, two students per classroom does not represent universal screening and may not reflect classroom changes or lack of changes in accuracy. The third research question examined the Total Behavior ratings of all students but did not include a measure of accuracy. Future studies should assess the accuracy of all teacher ratings to align with the purpose of universal screening. Furthermore, District A recruited secondary participants with SAEBRS scores generally well above risk cut-off (Total

Behavior score of 36), and thus there is less variability in District A SAEBRS scores and observations of behavior. The generalizability of the results is limited because demographic data was not collected for teacher and student participants. Gender for teacher and observation participants was estimated instead of given by participants. Based on estimations, teachers were predominately female and over half of observed students were male. The absence of demographic data limits the ability to conclusively determine the representativeness of the sample and add covariates to the multilevel models. Although individual demographic data was not collected, school-level demographics suggest that data was collected from both rural and suburban settings, a wide range of free and reduced price lunch eligibility, and populations with diverse as well as predominately White populations. School-level data should closely resemble Total Behavior scores because teachers rated their entire classrooms. However, the race/ethnicity, special education status, and socioeconomic status (as indicated by free and reduced price lunch eligibility) of the observed students is unknown and may relate to teacher perception of behavior (Mason et al., 2014). Further, teacher demographic characteristics may also impact ratings of behavior (Mashburn et al., 2006; Peters, Kranzler, Algina, Smith, & Daunic, 2014) and future research should collect detailed demographic data for both teachers and students.

In sum, as with most applied large-scale studies, there were a large number of limitations. However, the study applied a novel approach to a needed area of research. The results are not conclusive but highlight the need for future research on training teachers for universal screening measures.

Conclusion

The current study contributes to the literature by examining teacher training for universal SEB screening. The success of early intervention systems depends on accurate identification systems and the current study is one step forward to improving teacher accuracy. Although the current study did not find significant differences from FOR training, future research is warranted on the training method. Results were limited by the ability of observations to serve as an indicator of "true" behavior, ability to observe social behavior, and students selected for observations. Further, the null effect may be due to low power of teachers included in the observation analyses and relatively brief nature of the intensive training and lack of ongoing training. Examining the accuracy of universal rating scales in applied settings is difficult because there is a lack of feasible indicators of "true" behavior over one month. The current study can guide future research and establish procedures for an intensive teacher training for rating students with a universal screening measure.

References

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychological bulletin*, *101*, 213-232.

Achenbach, T. M., & Rescoria, L. A. (2001). *Manual for the ASEBA school-age forms and profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Aguinis, H., Mazurkiewicz, M. D., & Heggestad, E. D. (2009). Using web-based frame-of-reference training to decrease biases in personality-based job analysis: An experimental field study. *Personnel Psychology*, *62*, 405-438.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research.

Angkaw, A. C., Tran, G. Q., & Haaga, D. A. (2006). Effects of training intensity on observers' ratings of anxiety, social skills, and alcohol-specific coping skills. *Behaviour Research and Therapy*, *44*, 533-544.

Auger, R. W. (2004). The accuracy of teacher reports in the identification of middle school students with depressive symptomatology. *Psychology in the Schools*, *41*, 379-389.

Bergeron, R., Floyd, R. G., McCormack, A. C., & Farmer, W. L. (2008). The generalizability of

externalizing behavior composites and subscale scores across time, rater, and instrument. *School Psychology Review*, 37, 91–108.

Breslau, J., Lane, M., Sampson, N., & Kessler, R. C. (2008). Mental disorders and subsequent educational attainment in a US national sample. *Journal of Psychiatric Research, 42*, 708-716.

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

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*, 13-35.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist, 32*, 513-531.

Cairns, R. B., & Green, J. A. (1979). How to assess personality and social patterns: Observations or ratings. *The analysis of social interactions: Methods, issues, and illustrations*, 209-226.

Catalano, R. F., Berglund, M. L., Ryan, J. A., Lonczak, H. S., & Hawkins, J. D. (2002). Positive youth development in the United States: research findings on evaluations of positive youth development programs. *Prevention & Treatment*,*5*, 98-124.

Chafouleas, S. M. (2011). Direct Behavior Rating: A review of the issues and research in its development. *Education and Treatment of Children*, *34*, 575-591.

Chafouleas, S. M., Christ, T. J., Riley-Tillman, T. C., Briesch, A. M., & Chanese, J. A.

(2007). Generalizability and dependability of direct behavior ratings to assess social behavior of preschoolers. *School Psychology Review*, *36*, 63-79.

Chafouleas, S. M., Kilgus, S. P., Riley-Tillman, T. C., Jaffery, R., & Harrison, S. (2012). Preliminary evaluation of various training components on accuracy of Direct Behavior Ratings. *Journal of School Psychology*, *50*, 317-334.

Chafouleas, S. M., McDougal, J. L., Riley-Tillman, T. C., Panahon, C. J., & Hilt, A. M. (2005). What do daily behavior report cards (DBRCs) measure? An initial comparison of DBRCs with direct observation for off-task behavior. *Psychology in the Schools*, *42*, 669-676.

Chafouleas, S. M., Riley-Tillman, T. C., Jaffery, R., Miller, F. G., & Harrison, S. E. (2013). Preliminary investigation of the impact of a web-based module on Direct Behavior Rating accuracy. *School Mental Health*, *7*, 92-104.

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*, 201-213.

Christ, T. J., Riley-Tillman, T. C., Chafouleas, S., & Jaffery, R. (2011). Direct behavior rating: An evaluation of alternate definitions to assess classroom behaviors. *School Psychology Review*, *40*, 181-199.

Cohen, J. (1988). *Statistical power analysis of the behavioral sciences*. (2nd ed.). New York: Academic Press.

Cook, D. A., Dupras, D. M., Beckman, T. J., Thomas, K. G., & Pankratz, V. S. (2009). Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *Journal of General Internal Medicine*, *24*, 74-79.

Cook, C. R., Volpe, R. J., & Livanis, A. (2010). Constructing a roadmap for future universal screening research beyond academics. *Assessment for Effective Intervention*, *35,* 197-205.

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

Cunningham, J. M., & Suldo, S. M. (2014). Accuracy of teachers in identifying elementary school students who report at-risk levels of anxiety and depression. *School Mental Health*, *6*, 237-250.

De Los Reyes, A. (2013). Strategic objectives for improving understanding of informant discrepancies in developmental psychopathology research. *Development and psychopathology*, *25*, 669-682.

De Los Reyes, A., & Kazdin, A. E. (2006). Conceptualizing changes in behavior in intervention research: The range of possible changes model. *Psychological Review*, *113*, 554-583.

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin, 141*, 858-900.

Dirks, M. A., De Los Reyes, A., Briggs-Gowan, M., Cella, D., & Wakschlag, L. S. (2012). Annual research review: Embracing not erasing contextual variability in children's behavior–theory and utility in the selection and use of methods and informants in developmental psychopathology. *Journal of Child Psychology and Psychiatry*, *53*, 558-574.

Doyle, A., Ostrander, R., Skare, S., Crosby, R. D., & August, G. J. (1997). Convergent and

criterion-related validity of the behavior assessment system for children-parent rating scale. *Journal of Clinical Child Psychology*, *26*, 276-284.

DuPaul, G. J., Volpe, R. J., Jitendra, A. K., Lutz, J. G., Lorah, K. S., & Gruber, R. (2004). Elementary school students with AD/HD: Predictors of academic achievement. *Journal of School Psychology*, *42*, 285-301.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child development*, *82*, 405-432.

Essex, M. J., Kraemer, H. C., Slattery, M. J., Burk, L. R., Thomas Boyce, W., Woodward, H. R., & Kupfer, D. J. (2009). Screening for childhood mental health problems: Outcomes and early identification. *Journal of Child Psychology and Psychiatry*, *50*, 562-570.

Fabiano, G. A., Chafouleas, S. M., Weist, M. D., Sumi, W. C., & Humphrey, N. (2014). Methodology considerations in school mental health research. *School Mental Health*, *6*, 68-83.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology*, *45*, 117-135.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of child psychology and psychiatry*, *38*, 581-586.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*, 1337-1345.

Gresham, F. M., & Elliott, S.N. (2008). *Social Skills Improvement System: Rating Scales Manual*. Minneapolis, MN: Pearson Assessments.

Guerra, N. G., & Bradshaw, C. P. (2008). Linking the prevention of problem behaviors and positive youth development: Core competencies for positive youth development and risk prevention. *New Directions for Child and Adolescent Development, 122*, 1-17.

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

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, 92, 490 – 499.

Hightower, A. D., Work,W. C., Cowen, E. L., Lotczewski, B. S., Spinnell, A. P., Guare, J. C., et al. (1986). The teacher-rating scale: A brief objective measure of elementary children's school problem behaviors and competencies. *School Psychology Review*, *15*, 393-409.

Hinshaw, S. P., Han, S. S., Erhardt, D., & Huber, A. (1992). Internalizing and externalizing behavior problems in preschool children: Correspondence among parent and teacher ratings and behavior observations. *Journal of Clinical Child Psychology*, *21*, 143-150.

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

Hintze, J. M., Volpe, RJ, & Shapiro, ES (2008). Best practices in the systematic direct observation of student behavior. *Best Practices in School Psychology*, *4*, 993-1006.

Hosp, J. L., Howell, K. W., & Hosp, M. K. (2003). Characteristics of behavior rating scales implications for practice in assessment and behavioral support. *Journal of Positive Behavior Interventions*, *5*, 201-208.

Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer

ratings: A meta-analysis. *Psychological Methods*, *4*, 403-424.

Jones, E. E., & Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. E. Kanouse, H. H. Kelly, R. E. Nisbett, S. Valins, & Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 79–94). Morristown, NJ: General Learning Press.

Kamphaus, R. W., & Reynolds, C. R. (2007). *Behavior Assessment System for Children—Second Edition (BASC–2): Behavioral and Emotional Screening System (BESS)*. Bloomington, MN: Pearson.

Kessler Rc, (2005). Lifetime prevalence and age-of-onset distributions of dsm-iv disorders in the national comorbidity survey replication. *Archives of General Psychiatry, 62*, 593-602. doi: 10.1001/archpsyc.62.6.593

Kilgus, S. P., Eklund, K., von der Embse, N. P., Taylor, C. N., & Sims, W. A. (2016). Psychometric defensibility of the Social, Academic, and Emotional Behavior Risk Screener (SAEBRS) teacher rating scale and multiple gating procedure within elementary and middle school samples. *Journal of School Psychology*, *58*, 21-39.

Kolko, D. J., & Kazdin, A. E. (1993). Emotional/behavioral problems in clinic and nonclinic children: correspondence among child, parent and teacher reports. *Journal of Child Psychology and Psychiatry*, *34*, 991-1006.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Lane, K. L., & Menzies, H. M. (2011). *Systematic screenings of behavior to support instruction: From preschool to high school*. New York: Guilford Press.

Lane, K. L., Parks, R. J., Kalberg, J. R., & Carter, E. W. (2007). Systematic screening at the

middle school level score reliability and validity of the student risk screening scale. *Journal of Emotional and Behavioral Disorders*, *15*, 209-222.

LeBel, T. J., Kilgus, S. P., Briesch, A. M., & Chafouleas, S. M. (2010). The impact of training on the accuracy of teacher-completed Direct Behavior Ratings (DBRs). *Journal of Positive Behavior Interventions, 12,* 55–63.

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist*, *35*, 125-141. doi: 10.1207/S15326985EP3502_6

Levitt, J. M., Saka, N., Romanelli, L. H., & Hoagwood, K. (2007). Early identification of mental health problems in schools: The status of instrumentation. *Journal of School Psychology*, *45*, 163-191.

Lievens, F., & Sanchez, J. I. (2007). Can training improve the quality of inferences made by raters in competency modeling? A quasi-experiment. *Journal of Applied Psychology*, *92*, 812-819. doi:10.1037/0021-9010.92.3.812

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*, 54-71.

Mashburn, A. J., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2006). Teacher and classroom characteristics associated with teachers' ratings of prekindergartners' relationships and behaviors. *Journal of Psychoeducational Assessment*, *24*, 367-380.

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

Mattison, R. E., Bagnato, S. J., Mayes, S. D., & Felix, B. C. (1990). Reliability and validity of

teacher diagnostic ratings for children with behavioral and emotional disorders. *Journal of Psychoeducational Assessment*, *8*, 509-517.

McConaughy, S. H., Mattison, R. E., & Peterson, R. L. (1994). Behavioral/emotional problems of children with serious emotional disturbances and learning disabilities. *School Psychology Review, 23*, 81-89.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology*, *69*, 147- ???

Merrell, K. W. (2001). Assessment of children's social skills: Recent developments, best practices, and new directions. *Exceptionality*, *9*, 3-18.

Minor, L. (2013). Generalizability and dependability of brief behavior rating scales for social skills (Doctoral dissertation). Retrieved from http://etd.lsu.edu/docs/available/etd-10162013-234347/unrestricted/Minor_diss.pdf.

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, *74*, 619–624.

Missouri Department of Education. (2017). School Report Card. Retrieved on March 24, 2017 from https://mcds.dese.mo.gov/guidedinquiry/School%20Report%20Card/School%20Report%20Card.aspx.

Moor, S., Ann, M., Hester, M., Elisabeth, W. J., Robert, E., Robert, W., & Caroline, B. (2007). Improving the recognition of depression in adolescence: Can we teach the teachers?. *Journal of Adolescence*, *30*, 81-95.

Muris, P., Meesters, C., Eijkelenboom, A., & Vincken, M. (2004). The self-report version of the

Strengths and Difficulties Questionnaire: Its psychometric properties in 8-to 13-year-old non-clinical children. *British Journal of Clinical Psychology*, *43*, 437-448.

National Association of School Psychologists. (2009). *Appropriate behavioral, social, and emotional supports to meet the needs of all students* (position statement). Bethesda, MD: Author.

New Freedom Commission on Mental Health (2003). *Achieving the promise: Transforming mental health care in America. Final Report* DHHS Pub. Vol. SMA—3-3832. Rockville, MD: Author retrieved: November 15, 2009 from http://www.mentalhealthcommission.gov/reports/FinalReport/toc.html.

O'Connell, M. E., Boat, T., & Warner, K. E. (Eds.). (2009). *Preventing mental, emotional, and behavioral disorders among young people:: Progress and possibilities*. Washington, DC: National Academies Press.

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*, 452-465.

Reynolds, C. R., & Kamphaus, R. W. (1998). *Behavior assessment system for children: Manual,* Second Edition. Circle Pines, MN: American Guidance Service, Inc.

Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior Assessment System for Children, Second Edition*. Circle Pines, MN: American Guidance Service.

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, *85*, 370-395.

Romer, N., & Merrell, K. W. (2012). Temporal stability of strength-based assessments:

Test–retest reliability of student and teacher reports. *Assessment for Effective Intervention, 38,* 185-191.

Satcher, D. (2000). Mental health: A report of the Surgeon General: Executive summary. *Professional Psychology: Research and Practice*, *31*, 5-13.

Schlientz, M. D., Riley-Tillman, T. C., Briesch, A. M., Walcott, C. M., & Chafouleas, S. M. (2009). The impact of training on the accuracy of Direct Behavior Ratings (DBR). *School Psychology Quarterly*, *24*, 73-83.

Severson, H. H., Walker, H. M., Hope-Doolittle, J., Kratochwill, T. R., & Gresham, F. M. (2007). Proactive, early screening to detect behaviorally at-risk students: Issues, approaches, emerging innovations, and professional practices. *Journal of School Psychology*, *45*, 193-223.

Shapiro, E. S. (1996). *Academic skills problems: Direct assessment and intervention* (2nd ed.). New York: Guilford Publications.

Smith, S. R. (2007). Making sense of multiple Informants in child and adolescent psychopathology: A guide for clinicians. *Journal of Psychoeducational Assessment*, *25*, 139-149.

Spybrook, J., Raudenbush, S. W., Congdon, R., & Martinez, A. (2009). Optimal design for longitudinal and multilevel research: documentation for the "Optimal Design" software. From http://www.wtgrantfoundation.org/resources/overview/research_tools.

Spool, M. D. (1978). Training programs for observers of behavior: A review. *Personnel Psychology, 31,* 853-888.

Stamoulis, D. T., & Hauenstein, N. M. (1993). Rater training and rating accuracy: Training for

dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology*, *78*, 994.

Stichter, J. P., & Riley-Tillman, T. C. (2014). Considering systematic direct observation after a century of research: Commentary on the special issue. *Behavioral Disorders*, *39*, 245-247.

Suen, H., & Ary, D. (1989). *Analyzing quantitative observation data*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, *73*, 497-506.

Thomas, D. E., Bierman, K. L., & Powers, C. J. (2011). The influence of classroom aggression and classroom climate on aggressive–disruptive behavior. *Child development*, *82*, 751-757.

Tomb, M., & Hunter, L. (2004). Prevention of anxiety in children and adolescents in a school setting: The role of school-based practitioners. *Children & Schools*, *26*, 87-101.

Uggerslev, K. L., & Sulsky, L. M. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology*, *93*, 711-719.

Volpe, R. J., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in classroom settings: A review of seven coding schemes. *School Psychology Review*, *34*, 454-474.

von der Embse, N. P., Pendergast, L. L., Kilgus, S. P., & Eklund, K. R. (2016). Evaluating the

applied use of a mental health screener: Structural validity of the Social, Academic, and Emotional Behavior Risk Screener. *Psychological assessment*, *28*, 1265-1275.

Walker, H. M., Severson, H. H., & Feil, E. G. (2010). *Systematic Screening for Behavior Disorders (SSBD)*. Sopris West.

Weist, M. D., Rubin, M., Moore, E., Adelsheim, S., & Wrobel, G. (2007). Mental health screening in schools. *Journal of School Health*, *77*, 53-58.

Werthamer-Larsson, L., Kellam, S. and Wheeler, L. (1991). Effect of first-grade classroom environment on shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology*, *19*: 585–602. doi: 10.1007/BF00937993

Wilson, M. J., & Bullock, L. M. (1989). Psychometric characteristics of behavior rating scales: Definitions, problems, and solutions. *Behavioral Disorders*, *14*, 186-200.

Woehr, D. J. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology*, *79*, 525-534.

APPENDICES

APPENDIX A

Observation Form

Student Number: _____ Date: _____ Observer: _____ Classroom and Activity: _____ Time start: _____

	1 Child 1			1 Child 2			2 Child 1			2 Child 2			3 Child 1			3 Child 2			4 Child 1			4 Child 2			5 Child 1			5 Child 2			6 Child 1			6 Child 2		
AE	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X
PS	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X
DI	+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--	

	7 Child 1			7 Child 2			8 Child 1			8 Child 2			9 Child 1			9 Child 2			10 Child 1			10 Child 2			11 Child 1			11 Child 2			12 Child 1			12 Child 2		
AE	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X
PS	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X
DI	+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--	

	13 Child 1			13 Child 2			14 Child 1			14 Child 2			15 Child 1			15 Child 2			16 Child 1			16 Child 2			17 Child 1			17 Child 2			18 Child 1			18 Child 2		
AE	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X
PS	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X
DI	+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--	

	19 Child 1			19 Child 2			20 Child 1			20 Child 2			21 Child 1			21 Child 2			22 Child 1			22 Child 2			23 Child 1			23 Child 2			24 Child 1			24 Child 2		
AE	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X
PS	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X
DI	+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--	

	25 Child 1			25 Child 2			26 Child 1			26 Child 2			27 Child 1			27 Child 2			28 Child 1			28 Child 2			29 Child 1			29 Child 2			30 Child 1			30 Child 2		
AE	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X
PS	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X
DI	+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--	

	31 Child 1			31 Child 2			32 Child 1			32 Child 2			33 Child 1			33 Child 2			34 Child 1			34 Child 2			35 Child 1			35 Child 2			36 Child 1			36 Child 2		
AE	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X
PS	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X	+	--	X
DI	+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--		+	--	

Observe 15 minutes with 15 second intervals

+ = displayed the target behavior

-- = Given an opportunity but did not display the target behavior (e.g. unengaged, not disruptive, or not interacting with peers during available time).

X = No opportunity. For prosocial, if academic task demand without expectation to work with peers. For AE, if unstructured nonacademic time.

Academic Engagement (AE) = Actively or passively participating in the classroom activity. For example: writing, raising hand, answering a question, talking about a lesson or task demand, listening to the teacher, reading silently, or looking at instructional materials

Prosocial (PS) = Interacting with peer(s) in a positive or neutral way such as working cooperatively, joining task sharing, and/or talking/listening to peers related to classroom instruction. If unstructured nonacademic, student is polite and engaging with peers or teacher on appropriate topics and making appropriate responses such as not putting down others, turn taking, eye contact, and nonverbals.

Disruptive (DI) = Action that interrupts regular school or classroom activity or inappropriate interactions. For example: out of seat, playing with objects, acting aggressively, talking/yelling about things that are unrelated to classroom instruction, bossy comments, and/or defiance/noncompliance.

	37 Child 1	37 Child 2	38 Child 1	38 Child 2	39 Child 1	39 Child 2	40 Child 1	40 Child 2	41 Child 1	41 Child 2	42 Child 1	42 Child 2
AE	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X
PS	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X
DI	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --

	43 Child 1	43 Child 2	44 Child 1	44 Child 2	45 Child 1	45 Child 2	46 Child 1	46 Child 2	47 Child 1	47 Child 2	48 Child 1	48 Child 2
AE	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X
PS	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X
DI	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --

	49 Child 1	49 Child 2	50 Child 1	50 Child 2	51 Child 1	51 Child 2	52 Child 1	52 Child 2	53 Child 1	53 Child 2	54 Child 1	54 Child 2
AE	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X
PS	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X
DI	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --

	55 Child 1	55 Child 2	56 Child 1	56 Child 2	57 Child 1	57 Child 2	58 Child 1	58 Child 2	59 Child 1	59 Child 2	60 Child 1	60 Child 2
AE	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X
PS	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X
DI	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --

	61 Child 1	61 Child 2	62 Child 1	62 Child 2	63 Child 1	63 Child 2	64 Child 1	64 Child 2	65 Child 1	65 Child 2	66 Child 1	66 Child 2
AE	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X
PS	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X
DI	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --

	67 Child 1	67 Child 2	68 Child 1	68 Child 2	69 Child 1	69 Child 2	70 Child 1	70 Child 2	71 Child 1	71 Child 2	72 Child 1	72 Child 2
AE	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X
PS	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X	+ -- X
DI	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --	+ --

Observe 15 minutes with 15 second intervals

+ = displayed the target behavior

-- = Given an opportunity but did not display the target behavior (e.g. unengaged, not disruptive, or not interacting with peers during available time).

X = No opportunity. For prosocial, if academic task demand without expectation to work with peers. For AE, if unstructured nonacademic time.

AE = Actively or passively participating in the classroom activity. For example: writing, raising hand, answering a question, talking about a lesson or task demand, listening to the teacher, reading silently, or looking at instructional materials

PS = Interacting with peer(s) in a positive or neutral way such as working cooperatively, joining task sharing, and/or talking/listening related to classroom instruction. If it is unstructured nonacademic, the student is polite and engaging with peers or teacher on appropriate topics and making appropriate responses such as not putting down others, turn taking, appropriate eye contact, and nonverbals.

DI = Action that interrupts regular school or classroom activity or inappropriate interactions. For example: out of seat, playing with objects, acting aggressively, talking/yelling about things that are unrelated to classroom instruction, bossy comments, and/or defiance/noncompliance.

APPENDIX B

Familiarization Training

Universal Screening

Proactively identifies students

Prevention and early intervention more effective

Relies on teachers' objective evaluations of behaviors and frequencies

Students are selected for supports

Need for teachers to rate similarly

Universal Screening

Key Characteristics

- Standardized & Systematic
- Brief
- Periodic
- Key Indicators of Future Behavior

Teacher Role in Student Mental Health

Observer: Watch for signs and symptoms; Take notice in student's current skills in different domains of behavior

Catalyst: Conduct Universal Screening; Identify those at-risk & make a plan to address student need

Team Member: Follow-up, provide feedback on progress

Social, Academic, and Emotional Behavior Risk Screener (SAEBRS)

Brief Universal Screening Tool

- (19 items)

Technically Adequate

Evaluates student behavior in terms of overall general behavior and in narrow domains

Administration of SAEBRS

Total Behavior: 19 Items

- Social: 6 Items
- Academic: 6 Items
- Emotional: 7 Items

Rate the student's behavior on the following scale:

- 0 = Never, 1 = Sometimes, 2= Often, 3 = Almost Always

Ratings are based on frequency of behavior during the previous month

SAEBRS Scores

	At-Risk	Not At-Risk
Social Behavior	0-12	13-18
Academic Behavior	0-9	10-18
Emotional Behavior	0-17	18-21
Total Behavior	0-36	37-57

Brief Training Ends

INTENSIVE TRAINING CONTINUES

APPENDIX C

Frame of Reference Training

Brief Training Ends

INTENSIVE TRAINING CONTINUES

Social Behavior

Risk for Social Behavior Problems:

- Student displays behaviors that limit his/her ability to maintain age appropriate relationships with peers and adults

SAEBRS rates social skills and externalizing behaviors

Social skills:

- Cooperation with peers
- Polite and socially appropriate responses towards others

Externalizing behaviors

- Arguing
- Temper outbursts
- Disruptive behavior
- Impulsiveness

Academic Behavior

Risk for Academic Behavior Problems:

- Student displays behaviors that limit his/her ability to be prepared for, participate in and benefit from academic instruction

SAEBRS rates academic skills and academic behaviors

Academic Enablers:

- Interest in academic topics
- Prepared for instruction
- Produces acceptable work
- Academically engaged

Attention problems:

- Difficulty working independently
- Distractedness

Emotional Behavior

Risk for Emotional Behavior Problems:

- Student displays actions that limit his/her ability to regulate internal states, adapt to change and respond to stressful/challenging events

SAEBRS rates emotional competencies and internalizing emotional behaviors

Emotional Competencies:

- Adaptable to change
- Positive attitude

Internalizing Behaviors:

- Sadness
- Fearfulness
- Worry
- Withdrawal
- Difficulty rebounding from setbacks

Rating

Never
Less than once a week

Sometimes
1-2 times a week

Often
3-5 times a week

Almost Always
Several times a day

SAEBRS

Average Student

Think about an average student in the grade you teach. Think about their social, emotional, and academic behaviors and what you typically see from average students.

Complete the SAEBRS for this average student

Share your ratings with your table and come to consensus on what average looks like

Average Is....

Average is a range of ratings

Mostly "sometimes" on behavior problems

Mostly "often" on positive items

Majority of students will fall in this range [but not all]

Compare other students to "average"

Example Student: Social Behavior

Mary is a 1st grade student who argues with peers and adults several times a day, resulting in temper tantrums about once a week. During group activities, Mary cooperates with peers most days but disrupts other's work daily by talking out during carpet time and drumming loudly at her seat. Mary will occasionally say "please" or "thank you" and at least once a day interrupts others when they are talking.

Complete the social portion of the SAEBRS for Mary

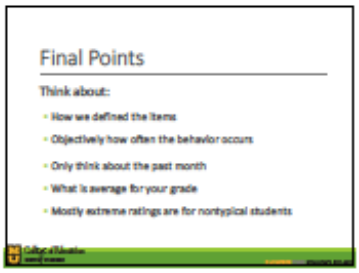
Mary's SAEBRS

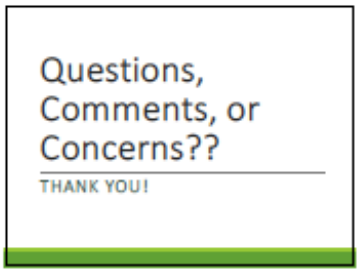
Share your ratings for Mary

Mary should be rated as:

- Arguing: Almost always
- Cooperation with peers: Often
- Temper Tantrums: Sometimes
- Disrupts others: Almost always
- Polite and socially appropriate responses towards others: Sometimes
- Impulsiveness: Almost always

- Do you think Mary would be risk for social behavior?


Final Points
Think about:
How we defined the items
Objectively how often the behavior occurs
Only think about the past month
What is average for your grade
Mostly extreme ratings are for nontypical students


Questions, Comments, or Concerns??
THANK YOU!

APPENDIX D

Intervention Fidelity Form

Intervention Fidelity Form

Name of Observer: _____

Date: _____

Component	Yes	No
Reviews screening purpose		
Discusses SAEBRS administration		
Reviewed select SAEBRS items		
Teachers rate an average student		
Discuss average ratings		
Trainer reviews how average should be rated and provides feedback		
Teachers rate vignette		
Teachers discuss vignette ratings		
Trainer reviews how example student should be rated and provides feedback		

Comments:

_____

_____

Time Start: _____ Time End: _____

VITA

Kristy Warmbold-Brann earned her Bachelor of Arts in Psychology with a minor in English in 2008 and Master of Arts in Counseling in 2011 from Truman State University. As a graduate student, Kristy worked as a tutoring program Site Coordinator, graduate research assistant, and graduate clinician. Kristy conducted research and consultation with school districts on initiating implementation of universal SEB screening with four school mental health initiatives. Kristy is finishing a pre-doctoral internship with Livingston County Special Services Unit where she provides school psychological services to rural school districts. Kristy frequently presents research and practitioner workshops at the National Association of School Psychologist Annual Convention. She is scheduled to earn her Doctorate of Philosophy in School Psychology from the University of Missouri in August 2017. Kristy has accepted a tenure-track Assistant Professor position in the School Psychology program at Miami University in Ohio starting August 2017.