Spring 2019

Characteristics of Behavior Rating Scales: Revisited

Ellen Cox
Western Kentucky University, ellen.cox329@topper.wku.edu

CHARACTERISTICS OF BEHAVIOR RATING SCALES: REVISITED

A Specialist Project
Presented to
The Faculty of the Department of Psychology
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
Of the Requirements for the Degree
Specialist in Education

By
Ellen Cox

May 2019

CHARACTERISTICS OF BEHAVIOR RATING SCALES: REVISITED

Date Recommended April 10, 2019

Carl Myers

Dr. Carl Myers, Director of Specialist Project

Dr. Sarah Ochs

Dr. Adam Lockwood

Cheryl A Davis 4/18/19

Dean, The Graduate School Date

ACKNOWLEDGMENTS

I would like to thank my specialist project director and professor, Dr. Carl Myers, for encouraging and helping me throughout this process and throughout my three years of graduate school, I could not have gotten this far without your constant support. I would also like to thank Dr. Sarah Ochs and Dr. Adam Lockwood for being on my committee and helping me make this specialist project the best it can be. Also, thank you to my friends, family, and classmates, your love and support over the past few years has meant the world to me.

CONTENTS

LIST OF TABLES

CHARACTERISTICS OF BEHAVIOR RATING SCALES: REVISITED

Ellen Cox May 2019 43 Pages

Directed by: Dr. Carl Myers, Dr. Sarah Ochs, and Dr. Adam Lockwood

Department of Psychology Western Kentucky University

This study was a replication of a study by Hosp et al. (2003), which looked at items on behavior rating scales to determine if they can be used to plan and monitor positive behavior interventions. For this study, ten forms of commonly used behavior rating scales were selected, and the *so what* and *dead man* tests were applied on each scale. Each item on the scale was placed into one of four categories: positive action, negative action, lack of positive action, and lack of negative action. Then, these categories were used to rate each scale to determine which subscales survived, or were deemed useful for measuring increases in positive behavior. Eight of the ten scales were found to contain a majority of negative action items and some lack of action items, neither of which are useful in measuring positive behaviors. Only two scales, the parent and teacher versions of the BERS-2, were found to contain all positive action items, and therefore were the only scales to fully survive the dead man test. The results of this study show that the majority of commonly used behavior rating scales today still do not contain primarily positive action items, and therefore have not majorly improved in the last fifteen years, although all of the behavior rating scales contained subscales that could have potential to plan and monitor positive behavior interventions.

Introduction

School psychologists are school team members that work with teachers, students, families, school staff, and the community by supporting academics, behavior, and mental health needs to create a safe and supportive school environment (National Association of School Psychologists [NASP], 2017). They do this through assessment, data collection, counseling, problem solving, intervention development, progress monitoring, and consultation, along with providing many other areas of support as needed, ranging from school-wide activities to the individual level (NASP, 2017).

School psychologists' roles in academic, cognitive, emotional, and behavioral assessment include planning and doing assessments, interpreting results, and engaging in data-based decision making to support students (NASP, 2016). They use a multifaceted approach, utilizing observations, record reviews, testing, and interviews when doing assessments (Fagan & Wise, 2007). Cognitive and achievement tests are widely used in most school assessments, while behavior rating scales, observations, and interviewing are typical for school psychologists to use as well (Shapiro & Heick, 2004). In fact, 60-90% of assessments conducted by school psychologists used interviews, observations, and rating scales to help make decisions (Shapiro & Heick, 2004).

Behavior rating scales used by school psychologists during evaluations need to be able to measure behaviors that can be used to guide interventions, and today many interventions are focusing on increasing positive behaviors (Walker, Cheney, Stage, & Blum, 2005). Therefore, the rating scales used must be able to measure children and adolescents' positive, observable behaviors. If behavior rating scales address positive

behaviors, they will be more useful for planning and monitoring positive interventions and goals.

This study is a replication of a study by Hosp, Howell, and Hosp (2003) in which the authors examined a variety of behavior rating scales to determine if they measured positive behaviors. In the current study, 10 forms of five behavior rating scales were reviewed, and the individual items on these protocols were analyzed to determine if they are action items, and if they are positive or negative. By analyzing the items for each scale, the following research questions were addressed:

1. Do current behavior rating scales primarily consist of positive action questions?

2. Based on the *so what* and *dead man* tests, as previously used by Hosp et al. (2003), are current behavior rating scales able to be used to plan and monitor positive interventions?

Literature Review

What are Behavior Rating Scales?

Behavior rating scales are instruments used by school and clinical psychologists to gain quantifiable information on an individual's behavior and social-emotional skills. The instruments typically list questions or phrases about behaviors that an individual may or may not present, to obtain information related to any behavior excesses or deficits. The rating scales are completed by an informant or multiple informants, such as an individual's parents and teachers, who have known the individual over an extended period of time (Campbell & Hammond, 2014).

Some authors distinguish between a behavior checklist and a behavior rating scale (Whitcomb, 2018). A behavior checklist only provides information on what behaviors are absent versus present. Often, behavior checklists are used to determine the number of diagnostic symptoms that are present related to a diagnosis (Hiller, Zaudig, & Mombour, 1990). Behavior rating scales indicate the degree of severity of the listed behaviors (Whitcomb, 2018). Behavior rating scales are widely used, especially in schools, because they are able to provide ratings and a standardized summary on various behavioral constructs related to a student's behavior indirectly and quickly (Whitcomb, 2018).

There are many different types of behavior rating scales available for use and they vary in multiple ways. They can vary with purpose, in that some behavior rating scales are used primarily for screening while others are better suited for classification or intervention monitoring purposes. Some rating scales, called broadband scales, are designed to cover a wide variety of behaviors to give a more comprehensive profile, while others, called narrowband scales, focus on fewer behaviors or those associated with

a specific disorder (Campbell & Hammond, 2014). For example, the ADHD Rating Scale-5 (ADHD-5) is a rating scale specifically designed to assess Attention-Deficit/Hyperactivity Disorder (ADHD) symptoms and progress monitor ADHD treatment, whereas the Behavior Assessment System for Children, Third Edition (BASC-3) is a popular comprehensive rating scale that assesses a wide variety of behaviors related to adaptive skills and externalizing and internalizing problems (Whitcomb, 2018).

Advantages of Behavior Rating Scales

There are many advantages to using behavior rating scales to gain information about an individual's behavior and social-emotional skills. Behavior rating scales are generally inexpensive, require little training to use, and collect information quickly, especially compared to other methods of data collection such as direct observations (Merrell, 2001). In addition, behavior rating scales can provide information on low-frequency behaviors, such as tantrums or violent behaviors that may not be seen by an outside observer conducting a direct observation (Whitcomb, 2018). Also, because scores on behavior rating scales are determined through normative data that are collected on individuals across the country of various ages, ethnicities, and disability statuses, this allows an individual's behavior to be compared to others in order to judge the severity of their behaviors (McConaughy & Ritter, 2014). This is important because it allows one to compare the student's behaviors to a representative group of students their age, not just students in the same school or community, to determine if a person's behaviors are age-appropriate, elevated, or lower than typical (Campbell & Hammond, 2014). Another important advantage of behavior rating scales is that they can provide information on students that the students are not able to provide themselves, because they are too young,

do not have adequate language skills, or are too disruptive or uncooperative (Merrell, 2001). Furthermore, behavior rating scales can quickly provide information on a wide variety of behaviors, and they are quantifiable scales that can be assessed in terms of reliability and validity (McConaughy & Ritter, 2014).

A significant advantage of behavior rating scales is that an individual's behavior can be compared through ratings from multiple informants in multiple natural environments (Whitcomb, 2018). Multiple informants, such as the student's teachers, parents, and even the students themselves, can rate behaviors using versions of behavior rating scales specifically designed for them to answer. For example, the BASC-3 has teacher, parent, and self-report forms (Reynolds & Kamphaus, 2015). This is an advantage because it allows "experts" on the students that are very familiar with them and their usual behaviors to rate them, instead of relying on an outside observer, such as the school psychologist, who may not observe a wide range of behaviors in a short period of time. Also, the informants rating the students based on their behavior in natural environments is an advantage because students' behaviors can vary drastically over time and across environments (Whitcomb, 2018). Having ratings from multiple environments, by the people who see them the most often in those environments, allows the school psychologist doing the assessment to get the best possible picture of how this student actually behaves in both school and home settings (Whitcomb, 2018).

Limitations of Behavior Rating Scales

There are also many limitations to using behavior rating scales. These limitations can be grouped into three general categories: bias, error variance, and questionable

appropriateness when used with culturally diverse students. These are not exclusive categories as there is some overlap in the limitations across categories.

Bias. Due to individual differences across people, informants' responses could potentially be too lenient, too harsh, or otherwise inaccurate due to intentional or unintentional bias (Chafouleas, Riley-Tillman, & Sugai, 2007; Whitcomb, 2018). For example, an informant could rate a student on a behavior rating scale in an overly positive way simply because the rater views him or her as a hard worker or as a nice person (halo effect). Similarly, an informant could tend to be overly severe (strictness bias) or generous (leniency bias) on ratings for all students, or an informant might tend to avoid endorsing any extreme ratings (e.g., "never" or "always"), no matter the item or student it is regarding (Whitcomb, 2018). All of these situations involve a form of bias when the informant is filling out the rating scales, and could cause inaccurate results.

In another form of bias, behavior rating scales can be susceptible to social desirability issues. This means that it can be difficult to determine if an informant is accurately answering the questions on the scale regarding the student, or if they are giving socially desirable answers (Merydith, Prout, & Blaha, 2003). Merydith et al. (2003) looked to see if the *Child Behavior Checklist/4-18* (*CBCL/4-18*) rating scale is susceptible to social desirability. In this study, the authors found that parents would sometimes respond differently to items about their child on the *CBCL/4-18* than they would on related items on one of the two social desirability scales used, showing that while their children did display the behaviors, they rated their child in a more positive light in order to be more socially desirable, especially when reporting on behaviors that

might be viewed negatively by others such as aggression or attention problems (Merydith et al., 2003).

Error variance. There are multiple types of error variance that could affect the validity of behavior rating scales, including rater, setting, time, and scale variance (Campbell & Hammond, 2014; Whitcomb, 2018). With rater variance, interrater agreement can be low when using behavior rating scales with multiple informants, especially between raters from different settings, such as between teachers and parents (Campbell & Hammond, 2014). The consistency of ratings from multiple informants about a child's behavior can often vary significantly, with parents, teachers, and children providing very different ratings (Mandal, Olmi, & Wilczynski, 1999). Situational factors, halo effects, and rater characteristics such as being too lenient or too strict are all factors to consider when looking at interrater agreement for a behavior rating scale (Mandal et al., 1999).

One study highlighted this issue by looking at parent and teacher agreement when diagnosing ADHD using the *Vanderbilt ADHD Diagnostic Teacher/Parent Rating Scales* (Wolraich et al., 2004). When looking to see parent-teacher agreement on which subtype of ADHD the child presented (Inattentive, Hyperactive/Impulsive, or Combined), they found poor interrater agreement between parent and teacher ratings on both the number of symptoms a child presented and if the child met ADHD criteria. These discrepancies are important, because if the diagnosis of ADHD for a child relies on multiple informants, discrepancies between raters would make this a difficult task or possibly make them ineligible (Wolraich et al., 2004). While school psychologists do not directly diagnose ADHD, they do rely on multiple informants to make special education

eligibility decisions and to design behavior intervention plans. Therefore, it is important to consider that just because ratings vary, this alone should not disqualify a child from receiving assistance (Wolraich et al., 2004).

Related to rater variance, setting variance can occur because behaviors that usually occur in one environment may not occur in another. For example, in school a child could exhibit certain behaviors due to factors or reinforcers that may not be present at home. Therefore, the student's teacher would rate that behavior as occurring often while the parents may not see the behavior at all and, as a result, provide low ratings of the same behavior (Whitcomb, 2018). Thus, different behaviors in different settings also contribute to poor interrater agreement and that can also make diagnostic decision-making difficult. However, it is also important to remember that a student's behaviors typically vary from setting to setting, so these differing ratings may actually be accurate (Whitcomb, 2018).

For time variance, the ratings of a student's behavior may change over time because the student's actual behavior changes over time or certain behaviors are more recent or salient to the rater. For example, one week a child may throw 20 tantrums and any behavioral items related to tantrums would likely be rated high, while the next week the child may only throw one or two tantrums, so those same items would likely be rated much lower (Whitcomb, 2018).

For scale variance, ratings on different scales designed to measure the same construct may differ due to the constructs being defined differently or through the use of different norming samples (Whitcomb, 2018). This can cause issues in the interpretation of results from multiple rating scales when different results occur for the same construct

being measured (Campbell & Hammond, 2014). For example, Myers, Bour, Sidebottom, Murphy, and Hakman (2010) examined scale variance between parent versions of the *BASC-2* and the *CBCL* preschool scales by controlling rater, setting, and time variance. Even when similarly named scales on the two instruments were significantly correlated, these researchers found that several corresponding scales from the two rating scales yielded significantly different mean scores. This means that different interpretations of results (i.e., average range vs. clinically significant) were obtained on the two rating scales measuring the same constructs. This can make deciding which rating scale to use and interpreting the results of both rating scales together difficult (Myers et al., 2010).

Cultural diversity issues. An additional issue with behavior rating scales is their use with culturally and ethnically diverse students. Across ethnic groups, many studies have been conducted to determine if disproportionally rating ethnic groups higher for certain disorders or behaviors is an issue with behavior rating scales. Reid, Casat, Norton, Anastopoulos, and Temple (2001) looked at the *IOWA Conners* and found that African American students are more likely to screen positive for ADHD than European American children. Also, they found an impact between rater-student ethnicity, meaning that African American teachers did not rate the African American students as high as the European American teachers did and vice versa (Reid et al., 2001). Another early study, Epstein, March, Conners, and Jackson (1998), looked at differences in externalizing behavior ratings on the *Conners Teacher Rating Scale* between European American and African American students. Those authors questioned cultural and ethnic bias with rating scales and found that teachers tended to rate the African American children higher on externalizing behaviors than European American children for both girls and boys,

although they could not say these differences were for sure due to ethnic bias or due to actual behavior differences (Epstein et al., 1998).

More recent studies on the topic show no bias or mixed results. For example, Mason, Gunersel, and Ney (2014) reviewed 13 studies on the topic and found mixed evidence for ethnic bias in ratings. Two of the strongest studies they looked at in terms of sampling and data collection methods found no evidence of bias, while another study they looked at found evidence of bias due to violated positive ethnic stereotypes. More specifically, Asian students with ADHD were rated more severely than European American or African American students exhibiting similar behaviors, due to positive ethnic stereotypes of Asian students being violated, causing increased teacher bias when rating the Asian students. In this review of studies, cultural bias was found in five studies, ethnic bias was found in six, and no bias was found in three (Mason et al., 2014). Overall, it can be hard to tell if behavior rating scales are susceptible to cultural diversity issues and racial bias, as the many studies that exist on the topic have widely varying results. Therefore, behavior rating scales should be used and interpreted with caution when used with ethnically or culturally diverse students.

Best Practice for Using Behavior Rating Scales

School psychologists use behavior rating scales for screening, assessment, intervention planning, and intervention monitoring purposes. They are used routinely for screening purposes as they cover a wide variety of behaviors, are inexpensive and quick to score, and have strong psychometric properties (Campbell & Hammond, 2014; McConaughy & Ritter, 2014). Behavior rating scales can be useful when designing interventions because they can help prioritize treatment targets and goals, and they can be

used to determine the location in which an intervention should take place, depending on the settings where problem behaviors are indicated to occur (Campbell & Hammond, 2014). They can also potentially be used to assess progress during and after interventions. Brief or shortened forms of behavior rating scales should be used when progress monitoring or looking at short-term outcomes, while the full length form of a scale can be used to look at long-term outcomes or to see if there are new or still-occurring problem behaviors to be addressed (Campbell & Hammond, 2014). When monitoring interventions, the use of behavior rating scales should ideally be combined with direct observations to ensure accurate progress monitoring and treatment fidelity (Campbell & Hammond, 2014).

When using behavior rating scales, it is recommended the data be obtained from multiple raters, in multiple environments, and using multiple scales if possible (McConaughy & Ritter, 2014). Campbell and Hammond (2014) suggest having two settings, with two raters, and two scales per rater, in addition to using self-report, interviews, and observations to make classification and intervention decisions.

Using Behavior Rating Scales for Intervention Planning and Monitoring

While it is considered best practices to use multiple types of assessment data along with behavior rating scales in order to make intervention decisions (Campbell & Hammond, 2014), behavior rating scales are increasingly being considered when planning interventions (Whitcomb, 2018). When looking at behavior rating scale data specifically, there are multiple ways the responses obtained can be useful to plan interventions. Whitcomb (2018) suggests using one of two strategies to link the data to specific interventions. The first strategy is the *Keystone Behavior Strategy* in which

clusters of responses that are linked to certain disorders or problems are identified, and then interventions that are known to be functionally linked to impacting that particular problem or disorder are selected for use (Whitcomb, 2018). The second possible strategy is the *Template-Matching Strategy*, in which data from referred children are compared to data obtained from children who have high functioning levels of the desired social behaviors, and the largest discrepancies between them would become the targets for intervention (Whitcomb, 2018). Overall, it is preferable to use behavior rating scale and direct assessment data to target specific areas for social skills training, rather than just identifying deficits and recommending generic social skills training (Whitcomb, 2018).

Sometimes, behavior rating scales are designed to be used in planning interventions directly. A study by Elliot, Gresham, Frank, and Beddow (2008) looked at the intervention validity, or the ability of a scale to directly lead to an intervention, of two behavior rating scales that only assessed social skills. The authors looked at rating scales they developed, the *Social Skills Rating System* (*SSRS*) and the *Social Skills Improvement System* (*SSIS*), to determine their connections to potential interventions (Elliot et al., 2008). The authors noted that while many behavior rating scales do not lead directly to interventions, the *SSIS*, a revision of the *SSRS*, is designed to connect the assessment results directly to potential interventions by determining areas of social skill problems, strengths, competing problem behaviors, and potential deficits in skill acquisition or performance (Elliot et al., 2008). The authors state that by including all of these data, the scale is potentially more useful in helping guide intervention planning decisions. However, they still suggest considering multiple assessments, with multiple raters, in multiple settings (Elliot et al., 2008). Whitcomb (2018) also noted the *SSIS* was designed

to be linked to intervention ideas specifically, so its use in addition to other sources of information when planning interventions for social or behavior problems could lead to better intervention results and effectiveness.

Whitcomb (2018) posits that behavior rating scales are being used more frequently for progress monitoring and summative evaluation of interventions. Behavior rating scales are thought to be useful for the summative evaluation of intervention efficacy, while using shortened versions may also be useful for monitoring progress on a weekly or daily basis (Whitcomb, 2018). There have been many studies using behavior rating scales as both summative intervention tools and progress monitoring tools (Gresham et al., 2010; McIntosh, Campbell, Carter, & Dickey, 2009; Volpe & Gadow, 2010).

As an example of a study that used a behavior rating scale as summative data, McIntosh et al. (2009) used the Behavior Assessment Scale for Children 2 (BASC-2) as a dependent variable along with office discipline referrals to measure an intervention's effects on ratings of problem and prosocial behavior and number of referrals. These authors found that the intervention (Check-in/Check-out) resulted in statistically significant improvements in ratings of problem behavior, prosocial behavior, and office discipline referrals for children with attention-maintained behavior, and they did not find significant differences for children with escape-maintained behavior (McIntosh et al., 2009). This study indicated that a behavior rating scale such as the BASC-2 could be useful to evaluate the effectiveness of an intervention (McIntosh et al., 2009).

Behavior rating scales, or selected items from behavior rating scales, have been recommended to monitor the effectiveness of interventions over shorter periods of time

as progress monitoring tools (Whitcomb, 2018). Behavior rating scales are typically too long to be used as a regular progress monitoring tool, so studies have sought to determine if shortened versions of behavior rating scales can be used to monitor interventions, and determine if they are still valid and reliable tools to measure behaviors and social skills (Gresham et al., 2010; Volpe & Gadow, 2010). Creating shorter rating scales with acceptable psychometrics enables the assessment of multiple constructs without asking too much of the informants (Volpe & Gadow, 2010). Gresham et al. (2010) looked at the teacher form of the Social Skills Rating System (SSRS) to determine how few items a brief behavior rating scale could contain while still adequately measuring social behaviors. They found that the optimal length for the brief rating scale for measuring externalizing behaviors while still being psychometrically valid and reliable was a 12-item scale, which would take about three minutes to complete (Gresham et al., 2010). Volpe and Gadow (2010) looked at shortened forms of the IOWA Conners Teacher Rating Scale and the Peer Conflict Scale and compared them to the full-length scales when looking at classroom inattention, over activity, aggression, and peer conflict. They found few significant differences between shortened and full-length scales, which support the use of abbreviated rating scales for progress monitoring (Volpe & Gadow, 2010).

Not all authors, however, support the use of behavior rating scales as progress monitoring tools. In a review of the literature on instruments used to measure behavior problems in children with autism, Hanratty et al. (2015) noted the BASC-2 and CBCL are often used to assess behaviors but reported a lack of evidence supporting they are sensitive to behavioral change. A sensitivity to behavioral change is necessary for a progress monitoring tool. Similarly, Wang, Sandall, Davis, and Thomas (2011)

determined that the Social Skills Rating Scale (SSRS) and Preschool and Kindergarten Behavior Scale (PKBS-2) are adequate instruments to measure social skills in young children with Autism Spectrum Disorder, but neither scale was found to be sensitive to change over time, and therefore they concluded that neither would be useful for progress monitoring or measuring the effects of intervention. For this reason they recommended that if behavior rating scales are used, other measures of behavior should be included to monitor intervention progress. As there are mixed results from various studies to whether behavior rating scales are able to show significant change due to interventions, they should be used in combination with other behavior monitoring methods when using them for progress monitoring purposes.

Positive Interventions

If using behavior rating scales to plan and monitor interventions, it is important to consider how and if they align with current intervention practices. Positive Behavior Interventions and Supports (PBIS) consist of widely popular practices in public schools today, and they focus on positively addressing discipline, academics, and social-emotional issues from a school-wide to an individual level (Walker et al., 2005). PBIS procedures are based on behavioral principles designed to be proactive and to teach alternative, appropriate behaviors. Furthermore, PBIS procedures are intended to replace more punitive, reactive procedures used in schools after behavior or social problems occur (Safran & Oswald, 2003). When the Individuals with Disabilities Education Act (IDEA) was revised in 1997, the federal special education law began requiring schools to use PBIS procedures for students in special education and for students whose behavior

can put them at risk for special education, so school psychologists need to be familiar with PBIS and its related procedures (Safran & Oswald, 2003).

PBIS procedures in schools typically involve an initial, school-wide screening to determine students who may be at-risk for behavioral or social problems, then implementing school-wide positive behavior supports through a response to intervention model, similar to those used for academics (Safran & Oswald, 2003). All students receive primary positive supports, through school-wide or classroom-level programs and disciplinary procedures. Additionally, school-wide supports are implemented in school settings such as hallways and cafeterias. Classroom and group-supports are implemented for specific groups of at-risk students who need additional supports. Individual students identified as high-risk through screeners and office discipline referrals and who are not making adequate progress through more universal supports receive individualized interventions (Safran & Oswald, 2003). As schools increase their focus on positive behaviors through implementing PBIS, it is important for school psychologists to consider not only if their interventions are positively based, but also if the rating scales they use are aligned with these practices as well (Bukley & Epstein, 2004).

At the tertiary level of PBIS, students who do not respond to primary and secondary interventions are receiving highly individualized interventions intended to stop problem behaviors, teach new skills, improve social functioning, and reinforce appropriate behaviors. It is especially important at this level to have appropriate tools to plan and progress monitor these interventions to make sure the student is making progress (Horner, Sugai, & Anderson, 2010). Planning and progress monitoring can also be important at the secondary level of PBIS, where students are receiving less intensive

interventions in a more small-group setting, but are still receiving interventions intended to teach positive behaviors and prosocial skills (Horner et al., 2010). Given PBIS focuses on fostering the growth of positive behaviors, not just decreasing the existence of negative ones, any behavior rating scales used in the process of planning interventions and monitoring behaviors targeted during PBIS secondary and tertiary interventions must be able to adequately measure if students are not only decreasing negative behaviors, but also increasing their levels of positive behaviors displayed (Hosp et al., 2003). In addition, using a strengths-based approach rather than a deficit-based approach can have many benefits that school psychologists should keep in mind, including increased parent involvement in the assessment process due to viewing a positive approach more favorably, and interventions and learning environments being able to build on and enhance students' strengths while helping students with various problem behaviors and deficits (Buckley & Epstein, 2004; Dinnebeil et al., 2013).

Purpose of this Study

Behavior rating scales can be used to screen for behavioral issues, inform educators of possible behaviors to target, and evaluate the outcome of interventions. Given the necessity for interventions to be positively-focused due to PBIS procedures in schools, there is a need to see if behavior rating scales can be used to measure positive behaviors. Hosp et al. (2003) looked at this issue, by examining items on 14 forms (e.g., parent, teacher) of 10 commonly used behavior rating scales to determine if they could be used to plan and monitor positive interventions. A list of those behavior rating scales is in Table 1. The authors stated that when using behavior rating scales to plan and implement positive interventions, it is important for the scale to address positive behaviors. If the

Table 1

Behavior Rating Scales Examined by Hosp et al. (2003)

Scales	Forms	Author(s)
Behavior Assessment Scales for Children (BASC)	Teacher, Parent, & Monitor	Reynolds & Kamphaus, 1992
Behavior Rating Profile-Second Edition (BRP-2)	Teacher & Parent	Brown & Hammill, 1990
Behavioral and Emotional Rating Scale (BERS)		Epstein & Sharma, 1998
Burks Behavior Rating Scales (Burks)		Burks, 1977
Child Behavior Checklist (CBCL)		Achenbach, 1991
Teacher's Report Form (TRF)		Achenbach, 2001
Revised Behavior Problem Checklist (RBPC)		Quay & Peterson, 1987
Social-Emotional Dimension Scale (SEDS)		Hutton & Roberts, 1986
Social Skills Rating System (SSRS)	Teacher & Parent	Gresham & Elliot, 1990
Walker-McConnell Scale of Social Competence and School Adjustment (Walker-McConnell)		Walker & McConnell, 1995

items on the scale focus on negative behaviors, alignment with the goals of the interventions can be lost, as decreases in negative behaviors does not always mean increases in positive behaviors (Hosp et al., 2003). The scales must focus on assessing positive behaviors that can replace negative behaviors in order to be useful for planning and monitoring positive intervention goals (Hosp et al., 2003).

In their study, Hosp et al. (2003) rated each item on the behavior rating scales to determine if they passed the *so what* and *dead man* tests. For the so what test, items were placed into one of four categories: positive action, negative action, lack of positive action, and lack of negative action. For the dead man test, items that were as lack-of-action items were reviewed to determine if an actual "dead man" would be given credit for the items (Hosp et al., 2003). See Table 2 for examples of items in each category.

Table 2

Examples of the So What and Dead Man Items

So What Test	Examples
Positive Action	Completes homework; plays with peers
Negative Action	Hits/bites; yells when angry
Lack of Positive Action	Does not complete work; does not participate
Lack of Negative Action	Does not hit; does not tease others
Dead Man Test	
Passes Test	Completes homework; hits others
Fails Test	Does not complete work; does not hit

After categorizing all items, Hosp et al. (2003) then totaled the percentages of each category for each of the behavior rating scales for the so what test, and they determined which subscales for each behavior rating scale survived or failed the dead man test to determine their overall usefulness for planning and monitoring positive interventions (Hosp et al., 2003). They found that while all of the scales consisted of a majority of action items, 10 of them were mostly negative action items, which means that 10 of the 14 forms of the instruments failed the so what test and did not align with the development of positive behaviors (Hosp et al., 2003). They also found that nine of the 14 tests had subscales that failed the dead man test, meaning that lethargy or even death could actually improve a student's score rather than the improvement being due to an increase in positive behavior (Hosp et al., 2003).

The results of this study are important, because they show that many rating scales commonly used in school psychology practice are potentially not useful for planning and monitoring positive interventions (Hosp et al., 2003). Positive interventions are becoming more and more relevant due to the increased use and requirement for PBIS in schools, and behavior rating scales are a widely used tool in schools. In order for behavior rating scales, or selected items from those scales, to be an appropriate tool to use for selecting target behaviors and monitoring positive interventions, they must be aligned with the goal of increasing positive behaviors.

More than 15 years ago, Hosp et al. (2003) found that the majority of commonly used behavior rating scales were not compatible for use with positive interventions. The purpose of this study is to replicate their methods and expand their study by evaluating behavior rating scales to determine if current versions can now be used to plan and

monitor positive interventions, and to see if behavior rating scales have generally increased in compatibility with measuring positive behaviors. Specifically, the following research questions will be addressed through this project:

1. Do current behavior rating scales primarily consist of positive action questions?

2. Based on the so what and dead man tests as previously used by Hosp et al. (2003), are current behavior rating scales able to be used to plan and monitor positive interventions?

For these two research questions, I hypothesize that:

1. Behavior rating scales have more positive action questions now than they found in the Hosp et al. (2003) study, as PBIS principles have become more popular.

2. If behavior rating scales are composed of mostly positive action items, then they would be able to adequately measure positive changes in a student's behavior.

Method

Materials

Ten forms of five behavior rating scales were identified for use in this study and are listed in Table 3. These behavior scales were chosen because they are either commonly used in practice today (Whitcomb, 2018), or they are updated versions of some of the scales used in Hosp et al. (2003). The first two scales listed in Table 3 are updated scales that Hosp et al. (2003) examined and the last three are additional scales.

As previously noted, broadband rating scales assess a wide variety of areas, such as internalizing and externalizing behavior problems and a wide variety of potential disorders, while narrowband instruments assess a more specific problem area, such as behaviors related to Autism Spectrum Disorder. Both were included in this study, as both

Table 3

Behavior Rating Scales Evaluated in the Current Study

Scales	Forms	Author(s)
Behavior Assessment System for Children, Third Edition (BASC-3)	Parent & Teacher	Kamphaus & Reynolds, 2015
Behavior and Emotional Rating Scale, Second Edition (BERS-2)	Parent & Teacher	Epstein, 2004
Social Responsiveness Scale, Second Edition (SRS-2)	Parent & Teacher	Constantino & Gruber, 2012
Autism Spectrum Rating Scales (ASRS)	Parent & Teacher	Goldstein & Naglieri, 2009
Conners, 3rd Edition (Conners-3)	Parent & Teacher	Conners, 2008

broadband and narrowband behavior rating scales can be used to plan and monitor behavior interventions (Whitcomb, 2018).

Many of these behavior rating scales include different forms for different age ranges. As such, the same gender and age that were used in the Hosp et al. (2003) study were used in this study to select the specific forms needed for a 10-year-old, fourth-grade male. They used these demographics because of the increased likelihood of a student to be diagnosed with Emotional Disturbance (ED) in late elementary school, and because most of the students who are identified with ED are male (Hosp et al., 2003).

Procedure

As this study is intended to be a replication of procedures used by Hosp et al. (2003), the procedure was identical to theirs for assessing whether these behavior rating scales are appropriate to use when planning and monitoring positive interventions. There are two different tests that were completed involving each behavior rating scale. The first test is called the so what test, which involves evaluating each item on each behavior rating scale protocol to determine if it addresses a positive or a negative behavior and if it involves an action or lack of action of the student. Therefore, each item on the protocol was placed into one of four categories by this thesis author: (a) positive action, (b) negative action, (c) lack of positive action, and (d) lack of negative action.

After every item on each behavior rating scale was placed into a category for the so what test by the first rater, a second rater independently repeated this process. The second rater for this study was a second-year school psychology graduate student that was given directions and examples to guide how to categorize each item on the scales.

Both raters' results were compared in order to determine the percent of inter-rater agreement. The total inter-rater agreement for all 10 scales was 91.3%, meaning the two raters placed 91.3% of the items in the same category. On the individual scales, the percent agreement ranged from 81.4% on the Conners 3 Teacher form to 100% on the BERS-2. Results are shown in Table 4. Generally, a minimum of 80% inter-rater agreement is deemed acceptable (Kennedy, 2005). Thus, the agreement rates in this study were at acceptable levels. Of the discrepancies that did occur between the two raters, all discrepancies were disagreements between two categories: Negative Action and Lack of

Table 4

Percent of Inter-rater Agreement for the So What Test

Scale	Percent Agreement
BASC-3 Parent	97.1%
BASC-3 Teacher	94.2%
BERS-2 Parent	100%
BERS-2 Teacher	100%
SRS-2 Parent	89.2%
SRS-2 Teacher	89.2%
ASRS Parent	90.1%
ASRS Teacher	90.1%
Conners 3 Parent	82.4%
Conners 3 Teacher	81.4%
Overall Inter-rater Agreement	91.3%

Positive Action. For example, for the item, "Has difficulty waiting for his/her turn," the two raters disagreed as to whether the statement was a Negative Action or a Lack of Positive Action. A third rater reviewed each discrepant item on each scale to determine which category the item should be placed. The third rater for this study was the faculty director of this thesis project and school psychology program coordinator that had knowledge of the different categories for the so what test and how to differentiate between them.

Using these final categories, the original rater scored each scale for the second level of evaluation, called the dead man test. The dead man test involves scoring each item according to whether or not an actual dead man could perform it. If an item was rated during the so what test as a positive or negative action item, it is considered to pass the dead man test as a literal dead man could not perform the action. These items are therefore given the minimum score on the scale (e.g., never, not true), while the lack of positive and negative action items are given the maximum score (e.g., always, very true) because a dead man could do them (Hosp et al., 2003). For example, "Does not turn in homework" would be rated as failing the dead man test, and therefore given the maximum score on the scale, because a dead man would not turn in his homework. The rationale behind the dead man test is that items must be observable and measureable in order to be a reliable measure of the increase in positive behaviors (Hosp et al., 2003). If scales have a mostly negative focus, any decreases in problem behavior can potentially cause improved behavior ratings. However, these improvements may only be due to sedation or the lack of any behavior occurring, and not due to increases in any positive, proactive behaviors (Hosp et al., 2003). As the scores for the dead man test are based on

the categories selected during the so what test, an inter-rater check was not necessary for the dead man portion of this study.

The scales were scored using the norms for a ten-year-old, 4th grade boy, the same age and gender used by Hosp et al. (2003). The reports for each scale were generated, and the results for each subscale on the behavior rating scales were compared to the minimum score required for the subscale to be considered clinically significant. For the BERS-2, a scaled score of 6 or less and a BERS-2 Strength Index score of 80 or below was needed to survive the dead man test, as these scores are in the High to Extremely High Emotional Behavioral Disorder probability ranges (Epstein, 2004). For all other rating scales, a T score of 60 or above is needed to survive, as these scores are considered to be elevated and above the average range (Conners, 2008; Constantino & Gruber, 2012; Goldstein & Naglieri, 2009; Kamphaus & Reynold, 2015). This allows a determination if *death*, or simply stopping negative behaviors but not increasing positive behaviors, could cause an improvement in subscale scores. In other words, assuming an original rating of clinically significant, the scores after *death* would no longer be clinically significant. This test will determine which subscales on each behavior rating scale survive and which ones fail the dead man test (Hosp et al., 2003).

To ensure the ratings and procedures in this study were completed in the same manner as in the original Hosp et al. (2003) study, the *Teachers' Report Form (TRF)* (Achenbach & Rescorla, 2001) was used to practice the so what and dead man tests and the results were compared to those found by Hosp et al. This particular behavior rating scale was chosen because this scale has not been updated since the Hosp et al. study was published. When looking at decision changes as a result of *death*, it was found that

Withdrawn and Internalizing were the only two subscales to survive the dead man test, which is the same result obtained by Hosp et al. (2003). Doing this ensured that the current researcher understood how to run the two tests and interpret them correctly in order to replicate the original study.

Results

The first research question sought to determine if current behavior rating scales primarily consist of positive action questions. The so what test was used to categorize each item on the behavior rating scales. For the so what test, the total number of items placed in each category (i.e., positive action, negative action, lack of positive action, lack of negative action) was recorded for every scale, and the percentages of items in each category for every scale was calculated. The results are presented in Table 5. For all 10 rating scales, the majority of the items were action items. However, for eight out of the 10 scales, the majority of the items were in the negative action category. This indicated that eight out of 10 of the rating scales in this study are able to measure reduction of negative behaviors, but not if they are being replaced by positive behaviors. Therefore, these eight scales fail the so what test. For two of the scales, both the Parent and Teacher forms of the BERS-2, all of the items were positive action items. Both forms of the BERS-2 survived the so what test, meaning these scales are able to measure increases in positive behaviors effectively.

Thus, the answer to the first research question is that many current behavior rating scales do not consist primarily of positive action questions. The BERS-2 was the only scale found to contain a majority of positive action items. This indicates that in the 15 years since the Hosp et al. (2003) study was published, many commonly used behavior rating scales are still focused on negative behaviors.

The second research question sought to determine if current behavior rating scales could be used to plan and monitor positive interventions based on the so what and dead man tests as previously used by Hosp et al. (2003). Thus, for the second part of this

Table 5

Percent of Items in Each Category on the Behavior Rating Scales for the So What Test

Rating Scale (number of items)	Positive Action	Negative Action	Lack of Positive Action	Lack of Negative Action
BASC-3 Parent (175)	30	66	4	0
BASC-3 Teacher (156)	34	60	6	0
Conners 3 Parent (108)	8	71	21	0
Conners 3 Teacher (113)	7	73	21	0
BERS-2 Parent (52)	100	0	0	0
BERS-2 Teacher (52)	100	0	0	0
SRS-2 Parent (65)	28	60	12	0
SRS-2 Teacher (65)	28	60	12	0
ASRS Parent (71)	31	58	11	0
ASRS Teacher (71)	31	58	11	0
Average of Percentages	40	50	10	0

Note. BASC-3 = Behavior Assessment System for Children, Third Edition; Conners 3 = Conners Third Edition; BERS-2 = Behavioral and Emotional Rating Scale, Second Edition; SRS-2 = Social Responsiveness Scale, Second Edition; ASRS = Autism Spectrum Rating Scales.

study, the dead man test, each scale was scored and compared to the cutoff scores for a significant result to determine if stopping all behavior (i.e., death) could cause a student's score to improve. In other words, a *lack of action*, instead of an increase in positive behavior, could cause a student's score on a scale to improve from the clinically significant range. Of the 10 rating scales, eight had scales or subscales that failed the dead man test. See Table 6 for a list of the scales for each instrument that survived or failed the dead man test. All 10 scales had scales or subscales that survived. The BERS-2 was the only rating scale that completely survived the dead man test, meaning none of the scores obtained on the BERS-2 would improve because of death.

Thus, in response to the second research question related to whether or not current behavior rating scales could be used to plan and monitor positive behavior interventions, generally, most instruments as a whole are still not designed to measure positive behaviors. Of the 10 scales evaluated in this study, the BERS-2 is the only behavior rating scale completely able to plan and monitor increases in positive behavior. The BERS was created to assess student's strengths and to be able to use student's strengths, not just their deficits, to design positive interventions (Buckley & Epstein, 2004). The BERS-2, now containing all positive action items, is able to do so even more. The BERS-2 is able to be used to measure student's strengths, use them to design interventions, and to monitor those interventions and changes in a student's strengths. The rest of the behavior rating scales did, however, contain subscales that survived the dead man test. Thus, specific subscales could be used to monitor positive behavioral interventions.

Table 6

Decision Changes as a Result of "Death"

	Subscales	
Rating Scale	Survived Dead Man Test	Failed Dead Man Test
BASC-3 Parent	Adaptive Skills Composite Adaptability Social Skills Leadership Functional Communication Activities of Daily Living Attention Problems Withdrawal	Externalizing Hyperactivity Aggression Conduct Problems Internalizing Anxiety Depression Somatization Aytypicality Behavior Symptoms Index
BASC-3 Teacher	Adaptive Skills Composite Adaptability Social Skills Leadership Study Skills Functional Communication Learning Problems School Problems Withdrawal	Externalizing Hyperactivity Aggression Conduct Problems Internalizing Anxiety Depression Somatization Attention Problems Atypicality Behavior Symptoms Index
Conners 3 Parent	Inattention Executive Functioning Learning Problems Peer Relations Inattentive Type	Hyperactivity/Impulsivity Aggression Conners 3 Global Index Hyperactive-Impulsive Type Conduct Disorder Oppositional Defiant

(continued)

	Subscales	
Rating Scale	Survived Dead Man Test	Failed Dead Man Test
Conners 3 Teacher	Learning Problems/Executive Functioning Total Executive Functioning Learning Problems Peer Relations Inattentive Type	Inattention Hyperactivity/Impulsive Defiance/Aggression Hyperactive/Impulsive Type Conduct Disorder Oppositional Defiant Conners 3 Total
BERS-2 Parent and Teacher (same scales and results)	Interpersonal Strength Family Involvement Intrapersonal Strength School Functioning Affective Strength BERS-2 Strength Index	
SRS-2 Parent and Teacher (same scales and results)	Awareness Cognition Communication Motor Social Communication and Interaction Total Score	Restricted Interests and Repetitive Behavior
ASRS Parent	Social/Communication DSM-5 Scale Peer Socialization Social/Emotional Reciprocity Attention	Unusual Behaviors Self-Regulation Total Score Adult Socialization Atypical Language Stereotypy Behavior Rigidity Sensory Sensitivity

(continued)

	Subscales	
Rating Scale	Survived Dead Man Test	Failed Dead Man Test
ASRS Teacher	Social/Communication	Unusual Behaviors
	DSM-5 Scale	Self-Regulation
	Peer Socialization	Adult Socialization
	Social/Emotional	Atypical Language
	Reciprocity	Stereotypy
	Attention	Behavioral Rigidity
	Total Score	Sensory Sensitivity

Note. BASC-3 = Behavior Assessment System for Children, Third Edition; Conners 3 = Conners Third Edition; BERS-2 = Behavioral and Emotional Rating Scale, Second Edition; SRS-2 = Social Responsiveness Scale, Second Edition; ASRS = Autism Spectrum Rating Scales.

Discussion

On all 10 of the behavior rating scales looked at in this study, the scales consisted of a majority action items rather than lack of action items. Action items, whether positive or negative, are able to give observable information about what a student is doing, while lack of action items are only able to indicate what they are not doing, which is not beneficial to plan or monitor interventions (Hosp et al., 2003). Eight out of 10 of the behavior rating scales looked at in this study had majority negative action items. This indicates that for the majority of the scales in this study, they are able to measure the presence and reduction of negative behaviors, but cannot provide adequate information on increases in positive behaviors. In other words, the eight scales with the majority of negative action items would not be able to distinguish between a student increasing positive behaviors and a student who simply stops performing, as ceasing performance can cause the student's scores to fall in the average range on certain subscales and overall scales.

The BERS-2, on the other hand, contained all positive action items, which are able to measure increases in a student's positive behaviors. This is an improvement from the results in Hosp et al. (2003), which found the original BERS contained 90% positive action and 10% lack of negative action items. The BASC-3 also had a small increase in its percentage of positive action items from the first edition of that behavior rating scale. In Hosp et al. (2003), the BASC Parent form had 25% positive action items, and the BASC Teacher form had 26%. The BASC-3 Parent form was found to have 30% positive action items, while the teacher form had 34%. However, both forms of the BASC-3 still

contained a majority of Negative Action items, and only a few subscales survived the dead man test.

The other scales in this study, the Conners 3, SRS-2, and ASRS, were not evaluated by Hosp et al. (2003). It was found in this study that they all contained a majority negative action items. Therefore, when using any of these scales to plan and monitor positive interventions, the treatment goals and subscales used to measure them would have to be carefully considered. For example, if the treatment goal for a student is to increase certain adaptive behaviors, the Adaptive Skills subscale on the BASC-3 may be useful as this subscale survived the dead man test. Likewise, if a student with autism has goals to increase social or communication skills, subscales on the SRS-2 or ASRS could be used.

Regardless of the behavior rating scales used to plan and progress monitor positive interventions, other measures of behavior (e.g., direct observation) should still be used to supplement any changes in behavior seen on the rating scales, to ensure any reductions in negative behaviors are being replaced by positive behaviors (Campbell & Hammond, 2014; Hosp et al., 2003). Since best practice for school psychologists is to use a multimodal, multireporter approach for any purpose, whether it be an evaluation or planning and monitoring an intervention, using other measures of behavior in addition to rating scales would be the most comprehensive and accurate data collection method (Campbell & Hammond, 2014). It is also important to consider the target behaviors of an intervention implemented, and if using a behavior rating scale would be the most appropriate method to monitor that behavior. For example, an intervention focusing on increasing social skills may be able to be adequately monitored with a scale or subscale

on a behavior rating scale, but an intervention focusing on a more specific behavior, such as increasing hand raising, may be monitored more effectively through other data-collection methods such as direct observations. Also, when considering using any behavior rating scale to plan and monitor an intervention it is important to consider other psychometrics of the scale, such as if it is sensitive to change and if it is designed for intervention planning and monitoring or more for classification and diagnostic purposes.

Limitations

A few issues were encountered during this study. The first issue was the difficulty discriminating between negative action items and lack of positive action items. As Hosp et al. (2003) states, these two categories are "generally different descriptions of the same concept" (p. 204). Therefore, it can be difficult to consistently rate them one way versus the other. The decision had to be made when placing test items into one of the four categories. When looking at inter-rater agreement, all of the discrepancies between raters one and two were between negative action and lack of positive action items. This rating difficulty impacted inter-rater agreement for many of the scales, especially the Conners 3 Parent and Teacher forms. The third rater looked at these discrepancies and decided based on the wording of the question if it was worded as a "lack of action" or an "action." For example, the third rater decided that all questions containing "Has trouble with…" implied a lack of action. This allowed a consistent rating of all items worded the same way in order to run the dead man test. However, some "lack of action" items, though worded as a lack of action, were not necessarily something a "dead man" would do. For example, on the Conners 3 an item is, "Has difficulty waiting for his/her turn." This was coded as a lack of positive action item, but a "dead man" would not necessarily have

difficulty waiting for his turn, which implies impulsiveness. Therefore, for many items it was difficult to determine if focus should be more on the wording to keep consistent ratings, or situational per question. If this process is repeated in future studies, clearer instructions on how to determine if items are negative action or lack of positive action would be necessary.

Another issue encountered in this study was many of the behavior rating scales used by Hosp et al. (2003) were not able to be replicated in this study. Some scales, such as the CBCL and TRF had not been updated since the original study. Other scales were not available for use or able to be scored for this study. Therefore, only two scales were truly able to be replicated, with three additional scales added to get a variety of commonly used behavior rating scales today. However, the methods used for this study and the research questions were the same as Hosp et al. (2003), so this study is still a replication of their study using current behavior rating scales to determine if they align with measuring positive behaviors.

Future Directions

The results of this study are important because they indicate that in the past 15 years, behavior rating scales have not evolved to emphasize positive behaviors to match the increasing focus on positive behavior interventions in schools. This could be because while behavior rating scales can be used in the intervention planning and monitoring process, they are also used frequently for evaluation and diagnostic purposes (Whitcomb, 2018). For evaluation or diagnostic purposes, the scales would just need to measure the presence or absence of certain behaviors, but they would not necessarily need to be positive behaviors. Many commonly used behavior rating scales today still show limited

usefulness for measuring increases in positive behaviors in students, just as they were not in Hosp et al. (2003). However, this study found that all behavior rating scales used in the study contained some subscales that are able to adequately measure positive behavior change. Therefore, these results can be used to select certain subscales that could be used for planning and monitoring positive behavior interventions, rather than using full behavior rating scales.

To expand on the current results, future research could examine additional rating scales to determine if other rating scales exsist that could be used with positive behavior interventions other than the BERS-2. Also, future studies could be completed using the subscales that were found to survive the dead man test to plan and monitor actual positive behavior interventions to determine the practicality and reliability of using behavior rating scales for this purpose. Future studies could help build on behavior rating scales' potential to be useful in planning and monitoring positive interventions, and potentially influence future revisions of behavior rating scales to be more aligned with measuring positive behavior change, by increasing the number of positive action items included on the scales.

References

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

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

Campbell, J. M., & Hammond, R. K. (2014). Best practices in rating scale assessment of children's behavior. In P. L. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (pp. 287-304). Bethesda, MD: National Association of School Psychologists.

Chafouleas, S., Riley-Tillman, T. C., & Sugai, G. (2007). *School-based behavioral assessment: Informing intervention and instruction.* New York, NY: Guilford Press.

Conners, K. C. (2008). *Conners 3rd edition*. Toronto, Ontario, Canada: Multi-Health Systems.

Constantino, J. N., & Gruber, C. P. (2012). *Social Responsiveness Scale, Second Edition*. Los Angeles, CA: Western Psychological Services.

Dinnebeil, L. A., Sawyer, B. E., Logan, J., Dynia, J. M., Cancio, E., & Justice, L. M. (2013). Influences on the congruences between parents' and teacher's ratings of young children's social skills and problem behaviors. *Early Childhood Research Quarterly, 28,* 144-152.

Elliot, S. N., Gresham, F. M., Frank, J. L., & Beddow III, P. A. (2008). Intervention validity of social behavior rating scales: Features of assessments that link results to treatment plans. *Assessment for Effective Intervention, 34,* 15-24. doi: 10.1177/1534508408314111

Epstein, M. H. (2004). *Behavioral and Emotional Rating Scale-2nd Edition: A strengths-based approach to assessment*. Austin, TX: PRO-ED.

Epstein, J. N., March, J. S., Conners, C. K., & Jackson, D. L. (1998). Racial differences on the Conners Teacher Rating Scale. *Journal of Abnormal Child Psychology, 26,* 109-118.

Fagan, T. K., & Wise, P. S. (2007). *School psychology past, present, and future* (3rd ed.). Bethesda, MD: National Association of School Psychologists.

Goldstein, S., & Naglieri, J. A. (2009). *ASRS: Autism Spectrum Rating Scales*. Toronto, Canada: Multi-Health Systems.

Gresham, F. M., Cook, C. R., Collins, T., Dart, E., Rasetshwane, K., Truelson, E., & Grant, S. (2010). Developing a change-sensitive brief behavior rating scale as a progress monitoring tool for social behavior: An example using the Social Skills Rating System - Teacher form. *School Psychology Review, 39,* 364-379.

Hanratty, J., Livingston, N., Robalino, S., Terwee, C., Glod, M., Oono, I., … McConachie, H. (2015). Systematic review of the measurement properties of tools used to measure behaviour problems in young children with autism. *PLoS ONE, 10,* 1-21.

Hiller, W., Zaudig, M., & Mombour, W. (1990). Development of diagnostic checklists for use in routine clinical care: A guideline designed to assess DSM-III-R diagnoses. *Archives of General Psychiatry, 47*, 782-784.

Horner, R. H., Sugai, G, & Anderson, C. M. (2010). Examining the evidence base for school-wide positive behavior support. *Focus on Exceptional Children, 42*, 1-14.

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.

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston, MA: Pearson.

Mandal, R. L., Olmi, D. J., & Wilczynski, S. M. (1999). Behavior rating scales: Concordance between multiple informants in the diagnosis of attention-deficit/hyperactivity disorder. *Journal of Attention Disorders, 3*, 97-103.

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.

McConaughy, S. H., & Ritter, D. R. (2014). Best practices in multimethod assessment of emotional and behavioral disorders. In P. L. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (pp. 367-389). Bethesda, MD: National Association of School Psychologists.

McIntosh, K., Campbell, A. L., Carter, D. R., & Dickey, C. R. (2009). Differential effects of a tier two behavior intervention based on function of problem behavior. *Journal of Positive Behavior Interventions, 11*, 82-93.

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

Merydith, S. P., Prout, H. T., & Blaha, J. (2003). Social desirability and behavior rating scales: An exploratory study with the Child Behavior Checklist/4-18. *Psychology in the Schools, 40*, 225-235. doi: 10.1002/pits.10077

Myers, C. L., Bour, J. L., Sidebottom, K. J., Murphy, S. B., & Hakman, M. (2010). Same constructs, different results: Examining the consistency of two behavior-rating scales with referred preschoolers. *Psychology in the Schools, 47*, 205-216. doi: 10.1002/pits.20465

National Association of School Psychologists. (2016). *NASP position statement: School psychologists' involvement in assessment*. Retrieved from https://www.nasponline.org/research-and-policy/professional-positions/position-statements

National Association of School Psychologists. (2017). *Who are school psychologists*? Retrieved from https://www.nasponline.org/about-school-psychology/who-are-school-psychologists

Reid, R., Casat, C. D., Norton, J., Anastopoulos, A. D., & Temple, E. P. (2001). Using behavior rating scales for ADHD across ethnic groups: The IOWA Conners. *Journal of Emotional and Behavioral Disorders, 9*, 210-218.

Reynolds, C. R., & Kamphaus, R. W. (2015). *Behavior Assessment System for Children manual* (3rd ed.). Bloomington, MN: Pearson.

Safran, S. P., & Oswald, K. (2003). Positive behavior supports: Can schools reshape disciplinary practices? *Exceptional Children, 69,* 361-373.

Shapiro, E. S., & Heick, P. F. (2004). School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. *Psychology in the Schools, 41,* 551-561. doi: 10.1002/pits.10176

Volpe, R. J., & Gadow, K. D. (2010). Creating abbreviated rating scales to monitor classroom inattention-overactivity, aggression, and peer conflict: Reliability, validity, and treatment sensitivity. *School Psychology Review, 39,* 350-363.

Walker, B., Cheney, D., Stage, S., & Blum, C. (2005). Schoolwide screening and positive behavior supports. *Journal of Positive Behavior Interventions, 7,* 194-204.

Wang, H. T., Sandall, S. R., Davis, C. A., & Thomas, C. J. (2011). Social skills assessment in young children with autism: A comparison evaluation of the SSRS and PKBS. *Journal of Autism and Developmental Disorders, 41,* 1487-1495.

Whitcomb, S. A. (2018). *Behavioral, social, and emotional assessment of children and adolescents* (5th ed.). New York, NY: Routledge.

Wolraich, M. L., Lambert, E. W., Bickman, L., Simmons, T., Doffing, M. A., & Worley, K. A. (2004). Assessing the impact of parent and teacher agreement on diagnosing Attention-Deficit Hyperactivity Disorder. *Developmental and Behavioral Pediatrics, 25,* 41-47.