Examining Rater Agreement After Changing the Response Format of the BRIEF

by

Bill Ryan Ferguson

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
May 9, 2011

Keywords: executive function, interrater reliability,
interrater agreement, rating scale, BRIEF

Approved by

Steven K. Shapiro, Chair, Associate Professor of Psychology
Jennifer M. Gillis, Assistant Professor of Psychology
Adrian Thomas, Professor of Psychology

Abstract

Deficits in Executive Function (EF) have been demonstrated in various types of childhood psychopathology. Current clinical practice encourages evaluating deficits in EF across multiple perspectives. The BRIEF is a parent- and teacher-completed rating scale designed to measure EF deficits in children. Previous studies have reported inconsistent interrater agreement (IRA) and interrater reliability (IRR) between parent and teachers. Differences in ratings may be influenced by the response format, begging the question whether a change in response format will improve IRA and IRR between parent- and teacher-completed BRIEF rating scales. Parents and teachers completed the BRIEF and BRIEF-R and mean differences, correlations, and Intraclass correlations (ICCs) were computed. Smaller mean differences and effect sizes between parent and teacher ratings revealed higher IRA for the BRIEF-R. There tended to be a slight rater preference for the BRIEF-R. Implications for rating differences and rater-reported preference of the BRIEF-R are discussed.

Acknowledgments

I would like to thank Steve Shapiro for his tireless support and mentorship, and sense of humor he maintained throughout this project. Furthermore, I thank Steve for the selfless devotion he clearly demonstrates with all of his students. I would also like to thank my committee members, Jen Gillis and Adrian Thomas, for their helpful input and feedback. I would also like to thank Jinyan Fan for his invaluable and gracious input regarding statistical procedures. To the Shapisk lab, I am so grateful to have the opportunity to interact with each of you on both personal and professional levels. I thank Andy Cohen for his expert advice and peer mentorship; and most importantly, his friendship. I also wish to thank Clarissa Mooney for her continuous, genuine warmth and encouragement. I thank Robert Butler, Melissa Cyperski, and Kristin Hiott for their comments, efforts, and support throughout the process. Thanks are due to individuals at the community agencies and school systems who provided their assistance in recruiting participants. Special thanks are due to Ottis Stephenson for his extraordinary efforts in recruiting participants in the Opelika City School system.

To my family, I would not be in the position I am today without their guidance and love. I especially want to thank my father for his wisdom and advice; my mother for her unconditional support; and my brother for keeping things entertaining. Lastly, and most importantly, I would like to thank my wife, Jenny, and my daughter, Ryan Lily, for teaching me more about life than any book or journal ever could. Their love is steadfast and their commitment is unwavering. I wish to express my everlasting gratitude for my two favorite girls.

Table of Contents

List of Tables

Introduction

Executive Function (EF)

Welsh and Pennington (1988) described executive functions (EFs) as neurocognitive processes that help to maintain an appropriate problem solving set to obtain a future goal. EFs are viewed as actions individuals perform to change behavior and consequently change their future (Smith, Barkley, & Shapiro, 2007). Furthermore, executive dysfunction and self-control appear to be directly affected by the functions of the frontal brain systems (i.e., prefrontal-striatal network) and its interconnections to the posterior and subcortical systems (e.g., cerebellum) (Smith et al., 2007; Welsh & Pennington, 1988).

A model proposed by Barkley (1997) links behavioral inhibition to four proposed executive functions (EFs): (a) nonverbal working memory (NVWM), (b) verbal working memory (VWM), (c) self-regulation of affect-motivation-arousal (SR), and (d) planning or reconstitution. The execution of these EFs is primarily dependent on behavioral inhibition. NVWM is best described as the ability to maintain and use nonverbal information (especially visual imagery) to control a motor response to obtain a goal. VWM is self-directed speech designed to assist in self-control, planning, and goal-directed behavior. Self- regulation of affect-motivation-arousal uses the skills discussed in the first two EFs (imagery and self-directed speech) to manipulate emotional states, which in turn govern the ability to induce the motivational states required for goal directed behavior. Lastly, planning is privatized self-directed behavior used for problem solving and goal-directed behavior (Smith et al., 2007). While not designed to specifically explore this model psychometrically, this study examines the

effects of response format, using a multi-dimensional EF rating form, on raters representing different situational contexts.

Attention-Deficit/Hyperactivity Disorder (ADHD)

Attention-Deficit/Hyperactivity Disorder (ADHD) is considered a neurobiological disorder that is commonly diagnosed in children who experience significant attention problems, and/or impulsivity and excessive activity. ADHD is characterized by two different dimensions of behavior: inattention-disorganization and hyperactivity-impulsivity (Barkley, 2006a). Furthermore, the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV-TR; American Psychiatric Association (APA), 2000) criteria require that symptoms be present prior to the age of 7, impairment is pervasive (i.e., impede the individual's ability to function in multiple settings; such as school and home), symptoms have lasted for at least 6 months, and are abnormal for what is expected at the individual's current developmental stage. ADHD consists of three subtypes: Predominantly Inattentive (ADHD-I), Predominantly Hyperactive-Impulsive (ADHD-H), and Combined (ADHD-C). In order to meet criteria for ADHD-I, the individual must present with at least 6 of 9 symptoms of inattention (e.g., *often has difficulty organizing tasks and activities, often does not seem to listen when spoken to, is often distracted by extraneous stimuli, etc.).* Individuals diagnosed with ADHD-H present with at least 6 of 9 symptoms of hyperactivity-impulsivity (e.g., *often fidgets, "driven by a motor," often has difficulty awaiting turn, etc.).* Finally, individuals who meet criteria for ADHD-C must meet criteria for both the Predominantly Inattentive Type and the Predominantly Hyperactive-Impulsive Type (APA, 2000).

Epidemiological studies suggest that 3% - 7% of the childhood population meet criteria for ADHD (APA, 2000). Given the core nature of the deficits defining ADHD, children who

carry this diagnosis often tend to experience problems in many facets of life such as: problematic peer relationships, conduct problems, and substance abuse (Bagwell, Molina, Pelham, & Hoza, 2001; Hinshaw, 1987; Molina & Pelham, 2003). Additionally, clinicians previously believed that most individuals would actually outgrow ADHD once they passed adolescence; however research suggests many individuals will be negatively impacted by ADHD symptoms into adulthood (Barkley, Fischer, Smallish, & Fletcher, 2002; Biederman, Faraone, & Miberger, et al., 1996). Approximately 66% of those diagnosed in childhood with ADHD will persist into adulthood (Barkley, 2006).

Due to the persistence in functional impairments, it is imperative that the professional mental health community continue to improve the understanding, conceptualization, assessment, and treatment of ADHD. Smith and colleagues (2006) state that decisions regarding the treatment of ADHD have not been guided by scientific theory; rather, treatment approaches were maintained if they worked and disregarded if they did not work. Smith et al. (2007) suggest that any plausible theory regarding the conceptualization of ADHD should "posit neuropsychological constructs related to the normal development of inhibition, self regulation, and executive function, and explain how they may go awry in ADHD" (p. 77).

Barkley (1997) proposes that ADHD can be best understood as a developmental delay in behavioral inhibition, which disrupts self-regulation. Behavioral inhibition is a multidimensional construct consisting of three interrelated processes. The first process involves inhibiting the prepotent response (immediate reinforcement is available or has been paired with this response), which is difficult for individuals with ADHD. The second response, and a vital component of self-regulation, is the ability to interrupt an ongoing response based on feedback. The third

inhibitory process is the ability to sustain goal-directed behavior in the face of distractions (i.e., freedom from distractibility), termed interference control (Barkley, 2006).

Recently, Willcutt and colleagues (2005) sought to determine the validity of the EF theory of ADHD. The authors analyzed 83 studies that used an EF measure to assess ADHD and concluded that the results clearly support the notion that EF weaknesses are significantly associated with ADHD. Moreover, "executive dysfunction in domains such as response inhibition, planning, vigilance, and working memory plays an important role in the complex neuropsychology of ADHD" (Willcutt et al., 2005; p. 1343). Even though additional research is still needed to determine exactly how EFs relate to ADHD, research generally supports their use to assist in the understanding of the disorder and important functional individual differences. However, it can be difficult to discern how the presence of ADHD affects a child's functioning when conducting the evaluation in an office setting.

Most of the clinical measures and rating scales used to assess ADHD do not focus specifically on EF. According to Smith et al. (2006) there is a divide between the assessment methods used in research and those used in practice. Many of the performance-based methods of assessment used in studies are not practical in the clinic setting, given the time and money required. The clinician may have a relatively short time frame in which to develop a clinical impression, whereas caregivers and teachers have a much larger behavioral sample to evaluate the child's executive dysfunction. Therefore, researchers suggest that ADHD evaluations should include data collection from multiple informants (e.g., parents and teachers) and across various settings (e.g., home and school) in order to develop a fuller understanding of the child's functioning (Smith et al., 2007). Consequently, parent- and teacher-completed rating scales are beneficial assessment tools. The current study explored interrater agreement while modifying the

response format of an EF rating scale. These modifications may improve the clinical utility of the measure as well as shed light on factors associated with rater consistency.

Behavior Rating Inventory of Executive Function (BRIEF)

One solution to the difficulties found in using performance-based (lab) assessments of EF is to instead use informant rating scales. As Mahone et al. (2002) point out, numerous rating scales assess a wide range of behavioral functioning, such as the Behavior Assessment System for Children (BASC; Reynolds & Kamphaus, 2004), Child Behavior Checklist (CBCL; Achenbach, 1991), and Conners' Parent Rating Scale (CPRS; Conners, 1997). While these scales provide broad-based ratings of behavioral, emotional, and adaptive functioning, few scales address the multidimensional nature of the EF construct. Consequently, Gioia, Isquith, Guy, and Kenworthy (2000) developed the Behavioral Rating Inventory of Executive Function (BRIEF), a parent- or teacher-completed norm-referenced questionnaire designed to assess multiple theoretical aspects of EF in children. The BRIEF is normed for use with children 5-18 years of age. It focuses more on executive abilities, which are commonly described as emotional, behavioral, and metacognitive skills rather than psychopathology or specific behavior problems (Donders, 2002). Although the BRIEF was intended to assess deficits in EF related to several disorders (e.g., Pervasive Developmental Disorder, Tourette's disorder, high functioning autism, reading disorder, frontal lobe lesions, and mental retardation), there has been a significant amount of focus on the BRIEF with children with ADHD. Out of the eight scales of the BRIEF (Initiate, Working Memory, Plan/Organize, Organization of Materials, Monitor, Inhibit, Shift, and Emotional Control), the Working Memory and Inhibit scales are most closely associated with the diagnostic criteria for ADHD. However, elevations in scales such as the Plan/Organize, Organization of Materials, and Shift scales are also common in children with ADHD (Gioia et

al., 2000). The BRIEF contains two global scales. The Behavioral Regulation Index (BRI) includes the Inhibit, Shift, and Emotional Control scales. The Metacognitive Index (MI) includes the Initiate, Working Memory, Plan/Organize, Organization of Materials, and Monitor scales.

There are several unique advantages to using the BRIEF as opposed to performance-based measures. First, the BRIEF is intended to assess EF in children by monitoring commonly occurring behaviors that fit under the EF umbrella (Gioia et al., 2000). Second, the BRIEF is comprised of 86 questions and only takes about 10 minutes to complete, which is considerably less time than that required to obtain a score from an expensive and time-consuming performance-based measure of EF such as the Delis-Kaplan Executive Function System (Delis, Kaplan, & Kramer, 2001). Third, the BRIEF was designed to assess problems with EF over a 6 month period, which may help to generalize assessment of behavioral phenomena and apply the chronicity criteria inherent in ADHD. Finally, the BRIEF facilitates gathering information from multiple informants across various settings, which as previously mentioned (Smith et al., 2007), is considered to generate valuable descriptive information regarding cross-situational impairments.

Despite the aforementioned benefits of using the BRIEF in the assessment process, several studies have documented an apparent discrepancy between parent- and teacher-completed scores on the BRIEF (Mares, McLuckie, Schwartz, & Saini, 2007; McCandless, & O'Laughlin, 2007; Sullivan, & Riccio, 2007). However, findings have been inconsistent. The discrepancy between parent- and teacher-completed scales varied across each study. The results from key studies will be discussed below.

Interrater Agreement

McCandless and O'Laughlin (2007) examined the diagnostic utility of the BRIEF in a clinical setting. The authors believed that the BRIEF would be a useful tool for both identifying children with ADHD and distinguishing between children with ADHD-I and ADHD-C. Previous research has found neuropsychological differences between children with ADHD-I versus children with ADHD-H or ADHD-C (Lockwood, Marcotte, & Stern, 2001). For example, in concordance with Barkley's model of ADHD, Gioia et al. (2007) hypothesized that difficulties with selective attention would be associated with lower scores on the Working Memory scale for children with ADHD-I. In addition to examining the clinical utility of the BRIEF, McCandless and O'Laughlin (2007) analyzed agreement between the parent and teacher reports as well as convergent validity between both reports and the BASC, which is a well established, norm-referenced, broad-band rating scale.

McCandless and O'Laughlin used correlational analyses to compare BRIEF scales to parent and teacher report on the BASC. These researchers found significant correlations between all of the parent-rated BRIEF scales and the Attention Problems and Hyperactivity scales on the BASC. In contrast, only six of eight teacher-rated BRIEF scales were correlated with the BASC Attention Problems scale, whereas all of the BRIEF scales were significantly correlated with the BASC Hyperactivity scale. Furthermore, "the overall agreement between parents and teachers on the BRIEF, as indicated by the global composite, was minimal ($r = .13$)" (McCandless & O'Laughlin, 2007; p.385). Correlations between parent and teacher ratings were significant on only three of the eight BRIEF scales (Inhibit, Plan/Organize, and Monitor).

McCandless and O'Laughlin also found significant group differences for parent and teacher ratings on the global BRI and MI scores. Parents rated children with ADHD-C

significantly higher than non-ADHD controls and children with ADHD-I on the BRI, and parents and teachers rated children with the ADHD-C higher on the MI than the non-ADHD controls. Compared to parent ratings, teachers rated working memory deficits much higher for the ADHD-I group. The authors concluded that the BRIEF was indeed capable of discriminating between ADHD and non-ADHD groups. A discriminant analysis classification revealed that 77.8% of the ADHD group was correctly classified, compared to 76% of the non-ADHD group. Cross-validated classification results determined that parent Inhibit scale and the teacher MI composite significantly predicted group membership for ADHD and non-ADHD subjects. Specifically, 77.1% of the original sample was correctly identified as either belonging to the ADHD group or the non-ADHD group. An additional discriminant analysis also revealed the BRIEF's ability to differentiate between ADHD diagnostic groups (i.e., ADHD-I, ADHD-C, non-ADHD) by correctly classifying 62.9% of the actual group membership (McCandless & O'Laughlin, 2007). These findings suggest that the BRIEF has clinical utility.

Similarly, Mares et al. (2007) used the BRIEF to compare parent and teacher reports of EF. The authors indicated that most studies using the BRIEF have used parent ratings and cited only two studies in which both the parent and teacher ratings were utilized with an ADHD sample. Mares et al. compared EF across the school and home environments, hypothesizing that teachers would report more EF deficits on the BRIEF due to the structured nature of the school setting. Additionally, the authors believed that the teachers' ratings would better predict symptoms of ADHD relative to parents' ratings.

Mares et al. (2007) found low levels of parent and teacher agreement on the BRIEF. The only scales reaching a statistically significant correlation were the Inhibit, Shift, Emotional Control, and Plan-Organize scales as well as the BRI composite. Inhibition was strongly

endorsed by parents and teachers, indicating that it is the greatest risk factor for ADHD. However, the correlations between the parent and teacher ratings were minimal (overall mean r = .16). Consequently, the authors concluded that there was little agreement between the parent- and teacher-completed scales on the BRIEF, commensurate with the findings of McCandless and O'Laughlin (2007). Teachers rated the children on all scales as having greater deficits of EF, which might suggest that teachers are either better able to identify EF impairments and/or children are actually showing more EF impairment at school than at home.

Finally, Sullivan and Riccio (2007) investigated parent and teacher BRIEF ratings across different groups. The authors compared children who met criteria for ADHD-I or -C, another clinical disorder (e.g., learning disorder, adjustment disorder, mood disorder, substance abuse disorders, and conduct disorder and oppositional defiant disorder), and children with no diagnosis. A significant group effect was found when using the parent ratings on the BRIEF with the three groups differing statistically on all but one scale (i.e., Organization of Materials). In contrast, a significant group effect was not found with the teacher BRIEF. A limited sample size for teacher BRIEFs was noted as a limitation.

Sullivan and Riccio (2007) also investigated diagnostic group differences of the parent and teacher forms of the Conners' Rating Scales – Revised (Conners' Parent Rating Scales; [CPRS] and Conners' Teacher Rating Scales; [CTRS]) short form (Conners, 1997). The study aimed to determine whether each scale could correctly predict group membership for children belonging to a no diagnosis group, ADHD group, and Other Clinical Diagnoses group. The authors found that both the BRIEF and Conners' scales were successful in distinguishing the ADHD group or other clinical diagnoses group from children without any diagnoses, but were not as successful in discriminating children with ADHD from children with other diagnoses.

Moderate yet significant correlations between parent and teacher ratings on the same BRIEF scales suggest a significant degree of agreement between the parents and teachers on the BRIEF. The parent and teacher ratings were also significantly correlated on the CPRS – Short Form and CTRS – Short Form. Sullivan and Riccio (2007) proposed that the moderate BRIEF correlations reflect the differences in perspective between raters as well as the variable nature of children's behavior across different settings (e.g., characteristics of the classroom or social situation).

Taken together, the literature has been inconsistent regarding the agreement between parents and teachers when using the BRIEF. Many researchers posit that the lack of interrater agreement between parents and teachers could be attributed to children's behavioral changes across the home and school environment (Mares et al., 2007; McCandless & O'Laughlin, 2007; Smith et al., 2007; Sullivan & Riccio, 2007). McCandless and O'Laughlin (2007) point out the fact that parents will likely observe, first-hand, the difficulties their children have with behavioral control. Teachers will undoubtedly have more opportunities to detect impairments in executive functioning associated with ADHD due to the academic demands placed on the child. In addition, teachers are able to obtain documented feedback from children (e.g., tests, homework, and coursework). It is not the purpose of this study to argue whether children's behavior differs across settings or informants provide different perspectives. This is well established, dating back to the seminal article by Achenbach, McConaughy, and Howell (1987). Rather the purpose of the present study is to determine if a change in the current response format of the BRIEF will have an effect on interrater agreement between parents and teachers and what this effect may reflect.

Scaling Issues

The BRIEF consists of a three-point Likert-type scale (Never, Sometimes, Often). One of the caveats associated with ratings is that the assumptions underlying the scale anchors are not always clear and therefore may be interpreted differently by separate observers (i.e., interpretation of the scale anchor by the teacher versus the parent). One person might interpret *almost always* to mean that the behavior occurs 99% to 100% of the time, whereas another person might interpret the same anchor to mean 90% to 100% of the time or even less (Sattler & Hoge, 2006). Parents may interpret *Often* as meaning that behavior X occurs nine times in the same day, while the teacher would not endorse *Often* unless she observed behavior X three times in the same day. Aiken and Groth-Marnat (2006) label this problem *ambiguity error,* or the failure to interpret items "correctly" due to the fact that the scale anchors are not explicit enough.

Kenney (1991) proposed six parameters that influence the level of agreement among raters:

> (1) Amount of information available to the judges. Kenney suggests that as judges are exposed to more of the child's behaviors, agreement will increase.
>
> (2) Extent to which the two judges observe the target behaviors at the same time. This parameter is important when raters (i.e., teachers and parents) observe the target behaviors during separate times of the day.
>
> (3) Degree to which different judges observe a child engage in a behavior and interpret it the same way. Two components of this parameter are particularly relevant to this proposed study. First, both the parent and the teacher have to agree that the child is engaging in the target behavior (e.g., child *is fidgety*); then they

need to accurately interpret the scale portions of the BRIEF (e.g., *Often* means 75% of the time to both the parent and teacher).

(4) Degree to which the child's behavior is consistent. That is, the child may act "*wild and out of control"* at home, but be well-mannered at school or vice-versa.

(5) Degree to which the judge's ratings are based on extraneous information or information not based on the child's behavior

(6) Amount of communication between raters (e.g., parent-teacher meetings)

Reid and Maag (1994) also commented on scaling issues, noting that it is still possible for two raters to have correlated scores even though they might not agree about the frequency of the child's behavior. For example, if the rating scales consisted of anchors such as *not at all, pretty much,* and *frequently,* and the first rater consistently rate the child *not at all* to *pretty much* while the second rater rated the child *pretty much* to *frequently* the scores would be highly correlated, but there would be a lack of rater agreement. As discussed above, it is possible, and perhaps likely, that the child's behavior will vary across multiple settings (e.g., school vs. home). However, Cairns and Green (1979; as cited in Reid & Maag, 1994, p. 346) suggest that agreement requires four assumptions to be met.

(1) Raters should have a common understanding of the attribute being rated.

(2) There should be a shared understanding of the typical behaviors that correspond to the attribute being rated.

(3) Raters should be able to accurately determine both the occurrence and nonoccurrence of behaviors related to the attribute being rated.

(4) Share a common metric in order to accurately scale the behaviors germane to the attribute being rated.

Reid and Maag (1994) suggest that agreement is affected by the difference in which raters interpret the amount of behavior associated with a frequency rating. Additionally, the authors note that standards and tolerance for the target behaviors can also impact the ratings. The teacher's idea of sloppy work may be vastly different from that of the parent's. Furthermore, the parent may have less difficulty in managing situations, whereas the opposite may be true for teachers (or vice-versa).

Anastasi and Urbina (1997) state that there are several ways to improve the accuracy of ratings. Similar to Sattler and Hoge (2006), they affirm that the underlying difficulty in rating scales involves the ambiguity of either trait names, scale units, or both. Anastasi and Urbina (1997) suggest that the ratings should be expressed in a manner that will be interpreted the same way by all raters. Namely, it is likely that the current anchors of the BRIEF (*Never, Sometimes, Often*) are interpreted differently across raters, thus affecting rater agreement between teachers and parents. That being said, it might prove useful to adjust the current anchors in order to decrease the amount of *ambiguity error* associated with the BRIEF (Aiken & Groth-Marnat, 2006).

With the above context in mind, studies comparing various response formats have failed to yield clear-cut advantages of using one over the other (Borman, 1979). However, developers of more recent rating scales maintain that there may be advantages of certain types of rating scales over others (Holland, Gimpel, & Merrell, 2001; Weathers, Newman, Blake, Nagy, et al., 2004). Merrell (2008) stresses that ratings are more accurate when there is a concrete and clear definition for each quality level, meaning that anchor points need to be clearly defined and meaningful to each rater. It is important to use the fewest number of rating levels possible, in addition to analyzing each anchor point in order to be certain they are useful in reliably

discriminating among the ratings. Although the traditional rating scale format (e.g., *Never, Sometimes, Often*) is flawed, its use continues. Merrell (2008, p 103) suggests an alternative to the traditional model, in which the "anchors to specified numerical rating points are very specifically connected to the estimated frequency of specified behaviors." He refers to this alternative format as the *frequency of behavior* format. In contrast to traditional formats, the *frequency of behavior* format ties the anchors and rating points to specific time periods (e.g., 0 = *Behavior does not occur/ No knowledge of behavior*, 1 = *Behavior occurs one to several times a month*, 2 = *Behavior occurs one to several times a week*, 3 = *Behavior occurs one to several times a day*, 4 = *Behavior occurs one to several times an hour*).

In fact, Holland et al. (2001) chose this rating format in developing the ADHD Symptoms Rating Scale (ADHD - SRS). This scale has shown adequate convergent validity and reliability (Holland, Gimpel, & Merrell, 1998). The authors of the ADHD – SRS believed that the use of the *frequency of behavior* format would lead to higher reliability and ultimately more precise ratings (Merrell, 2008). After parents and teachers completed the ADHD – SRS using both the traditional format and the *frequency of behavior* format, statistical analyses revealed that the accuracy of ratings and reliability were similar. Although this finding did not support their hypothesis, Holland et al. (2001) reported that both the teachers and the parents preferred to keep the *frequency of behavior* over the traditional format. Consequently, the *frequency of behavior* format was chosen under the premise that parent and teacher preference of this format denotes that it has user acceptability and social validity (Merrell, 2008).

Likewise, Blake, Weathers, Nagy, Kaloupek, et al. (1995), developed a frequency of symptoms scale for the Clinician-Administered PTSD Scale (CAPS), similar to the frequency of behavior format used in the ADHD - SRS. The CAPS is a psychometrically sound interview

considered to be the "gold standard" for the assessment of Posttraumatic Stress Disorder (PTSD) (Weathers et al., 2004). The CAPS has been used extensively in research and has well established and strong psychometric properties. More specifically, the authors hypothesized that the use of explicit anchors would improve the scale's psychometric characteristics. There are two different types of frequency prompts and rating scale anchors employed in the CAPS. The first type of frequency prompt (How often) is used for discrete symptoms, and has corresponding scale anchors (0 = *never*, 1 = *once or twice*, 2 = *once or twice a week*, 3 = *several times a week*, 4 = *daily or almost every day*). The second type of frequency prompt (How much of the time) is used for continuous symptoms, and has corresponding scale anchors (0 = *none of the time*, 1 = *very little of the time (less than 10%)*, 2 = *some of the time (approximately 20% - 30%)*, 3 = *much of the time (approximately 50% - 60%)*, 4 = *most of the time (more than 80%*)).

Regardless of the differences between the two measures (ADHD – SRS and CAPS), both sets of authors chose to adopt rating scale formats based on the frequency of specific behaviors or symptoms. The literature has neither clearly supported nor refuted the decision to employ a frequency component to these scales. Yet, authors continue to note the impact of the lack of specificity of anchors on a scale's ability to detect changes brought about by treatment (McMahon & Frick, 2007). Blake et al. (1995) noted that the decision to implement the frequency format was based on clinical rationale, but rather what seemed to make the most sense given the nature of the question. That is, the authors did not report basing their decision to use the frequency format on an empirically derived process. Additionally, despite little evidence of psychometric improvement the *frequency of behavior* format was supported by the teacher and parent preference. The psychometric properties of both the ADHD – SRS and the CAPS have been well established in the literature (Holland, et al., 1998; Weathers et al., 2004), suggesting

that further investigation into this scaling format for other constructs is warranted (Merrell, 2008).

Focus of the Current Study

The current study examined whether a change in response format will influence the rater agreement between parents and teachers. This study did not attempt to investigate whether a child's executive functioning is the same across multiple settings and raters. Rather, the goal was to explore whether a change in the current response format of the BRIEF affects the overall agreement between parent and teacher ratings on the BRIEF.

We predicted that changing the current response format on the BRIEF to a frequency-based format would increase interrater agreement and interrater consistency or reliability. Such results would influence test developers to consider the use of alternate rating scale response formats in future scales. 1) Changing the current response format from *Never*, *Sometimes*, and *Often* to either a continuous or discrete symptom response format would result in a higher level of agreement. We expected the parent and teacher ratings on the revised BRIEF to reach a statistically significant higher rate of agreement than the parent and teacher ratings on the original BRIEF. 2) Changing the response format would significantly influence the level of interrater reliability. That is, we compared the parent and teacher ratings on the original BRIEF to parent and teacher ratings on the revised BRIEF to explore whether the consistency of ratings between the parents and teacher improved.

Method

Participants

Participants were recruited from the community in East Central Alabama by publicizing through local businesses, gyms, psychologist's and physician's offices, and medical clinics. Further recruitment of participants occurred through local school systems and the Auburn University Psychological Services Center. Although the BRIEF was normed for use with youth ages 5 to 18, only parents of children in grades 1 through 5 were used in order to control somewhat for developmental and moreso, school context differences. Additionally, Achenbach et al. (1987) found higher overall correlations for ratings of 6-11-year-olds in their study, suggesting that children may be more cross-situationally consistent than adolescents, which further supports the use of children in this age range.

Thirty-seven parents were screened for inclusion into the study by using the demographic questionnaire (see Appendix A). Inclusion criteria for the study were that: (a) one parent and one teacher complete both the original and revised version of the BRIEF; and (b) the teacher have a minimum of one month of daily contact with the child being rated (Gioia, et al., 2000). Parents of children were excluded from the study if the child began taking psychoactive medication during the 1-month window, or if their current medication dosage was adjusted during that time. Additionally, we collected information about the length of time the teacher had been with the student as well as any pertinent classrooms changes so as to limit the effects of any

extraneous variables. Similarly, we inquired about any changes in marital status and/or major home environment changes.

Twenty-eight individuals met study criteria, but only 20 individuals (17 mothers, 2 fathers, and one legal guardian) completed the study. The primary reason for study non-completers was failure to return rating scales. Most parents were excluded from the study because their child was being homeschooled. No parent indicated that their child was on a medication for psychological issues or has had any recent changes in his or her child's medication status. Table 1 summarizes the demographic information for the 20 children who were rated in the study. The average age of the child was 8.7 (SD = 1.17), and the majority of children were female (60%) and Caucasian (75%). Data were collected from teachers representing eight different schools in the surrounding area.

Table 1

Demographic Characteristics of Children Being Rated

Demographic Variable	Frequency	%	χ^2
Gender			0.80
Female	12	60	
Male	8	40	
Race/Ethnicity			16.30*
African American	4	20	
White	15	75	
Other	1	5	
Grade			3.50
1st	2	10	
2nd	5	25	
3rd	6	30	
4th	5	25	
5th	2	10	
Total	20	100	

Note: Significant values indicate an unequal representation of demographic characteristics in this sample.
*p < .001

Measures

BRIEF. The BRIEF is a rating scale designed to measure executive function in youth 5 to 18 years of age. The psychometric properties of the rating scale are described in the manual (Gioia et al., 2000). Additional validity studies were reviewed earlier in this document. The BRIEF consists of two forms: the BRIEF – Parent Form and the BRIEF – Teacher Form. Both forms contain 86 items grouped into eight clinical scales, and three indexes. The eight scales on the BRIEF include: Inhibit, Shift, Emotional Control, Initiate, Working Memory, Plan/Organize, Organization of Materials, and Monitor. The raw scores for Inhibit, Shift, and Emotional Control combine to form the Behavioral Regulation Index (BRI). The raw scores for Initiate, Working Memory, Plan/Organize, Organization of Materials, and Monitor combine to form the Metacognition Index (MI). The overall score (Global Executive Composite; GEC) is comprised of all eight clinical scales. Raw scores on each of the scales and indexes are converted to T-scores based on norms obtained from a sample of 1,419 children and adolescents on the BRIEF – Parent Form and the sample of 720 children and adolescents on the BRIEF – Teacher Form. Internal consistencies (Cronbach's alpha) for the parent and teacher scales ranged from .80 – .98 with clinical and normative samples. The current scale employs a Likert scale format with three anchors (N = *Never*, S = *Sometimes*, O = *Often*).

BRIEF-Revised. The BRIEF-Revised (BRIEF-R) is a revision of the current form of the BRIEF, designed for the proposed study. The items are identical to the published scale, but the scale anchors were changed into one of two types of frequency formats. The first format is a discrete symptom format, requesting the informant to estimate a behavioral count (e.g., 0 = *never*, 1 = *once or twice*, 2 = *once or twice a week*, 3 = *several times a week*, or 4 = *daily or almost every day*). With discrete symptoms, the rater can typically distinguish both the initiation

and the cessation of the behavior of interest. The question to ask when assessing the frequency of a discrete symptom is *how often* behavior Y occurs.

The second type of frequency format is called the continuous symptom format and is most appropriate for symptoms that may not have a discrete beginning and end. Unlike the behavioral counting approach employed with discrete symptoms, the continuous symptom format measures the percentage of time behavior Y occurs. The question to ask when assessing the frequency of a continuous symptom is *how much of the time* Y occurs. Typical anchors may be 0 = *none of the time*, 1 = *very little of the time (<10%)*, 2 = *some of the time (approx. 20-30%)*, 3 = *much of the time (approx. 50-60%)*, or 4 = *most of the time (>80%)*. Weathers et al. (2004) used these two types of frequency formats in the CAPS. For the current study a panel comprised of 3 doctoral-level clinical psychology graduate students and one licensed (clinical) psychologist reviewed each BRIEF item to assess whether or not it would best fit the discrete or continuous symptom format. Each rater submitted a document labeling each item as either better fitting the continuous or discrete symptom format. Items that did not reach absolute agreement were further discussed by the panel in an open format. Members of the panel provided a rationale for their selected symptom format of each item; a discussion of each item ensued until unanimous agreement was obtained.

Procedure

Prior to completing either rating scale, parents completed a screening packet containing an informed consent form, a demographic sheet, a contact information sheet, and medication status form. Once consent was obtained, each respective teacher was contacted separately in order to obtain his or her consent to participate in the study. Consistent with IRB procedures, following school administration approval, the teacher was contacted by phone or email to assess

the level of interest in participating in the study. Once a teacher expressed an interest in participating in the study, a consent form was mailed to his or her school in order to be completed and returned in the enclosed self-addressed mailing envelope. After one parent and one teacher had returned the consent form, and it had been determined that both met study criteria, both individuals were mailed a rating scale.

The BRIEF or the BRIEF-R was given to the dyad for each child using a counterbalanced order of administration. The parents and teachers were instructed to complete the scale and return it using an enclosed addressed envelope. Both the teacher and parent were asked to refrain from discussing the scale with each other. Two weeks after the completed scales were received from both the parent and teacher, each informant was asked to complete and return the second scale and indicate which rating scale version they preferred.

Parents and teachers received a $10 monetary reward for completing the study. In addition, all participants who completed an initial screening packet were eligible to win one of two $25 cash prizes (one for teachers and one for parents).

Results

Tests for homoscedasticity and skewness were conducted before rating comparisons were pursued. These distribution statistics were generally within normal limits. Nonparametric statistics were computed due to the potential violations of assumptions held by parametric statistics. Specifically, the relatively small sample size decreased the likelihood of having a normal distribution. However, the nonparametric statistics (see Appendix C) resulted in virtually identical findings as the parametric statistics discussed below. Thus, the more powerful analyses were selected. Furthermore, to confirm the intended effect of counterbalancing the administration of version types, independent samples t tests were conducted. No order differences were found among raters or versions

Prior to making any comparisons of scores within raters (e.g., parent ratings on the BRIEF vs. parent on the BRIEF-R) or between raters (e.g., parent ratings on the BRIEF-R vs. teacher ratings on the BRIEF-R), it is important discuss the distinction between interrater reliability (IRR) and interrater agreement (IRA). Both IRR and IRA measure useful, but separate qualities of a particular set of ratings. A high level of IRR between parents and teachers would indicate that both informants are rank-ordering children in a similar or consistent fashion. Whereas IRR is typically assessed using correlational indices and is primarily concerned with the relative order of the ratings, IRA measures the absolute difference between the ratings on a particular variable and is sensitive to mean differences between judges (Tinsley & Weiss, 2000). For the present study, high IRAs would suggest that parent and teacher scores are very similar to

identical. It is crucial to understand that two judges can reach high IRR, but have virtually no IRA. Further discussions of the unique characteristics of IRR and IRA are available in the literature (LeBreton & Senter, 2008; Tinsley & Weiss, 2000).

Mean Differences (IRA)

Mean differences between the average parent and teacher rating were evaluated in order to determine the level of IRA. Prior to evaluating the differences between parent and teacher ratings, the average rating per scale was computed for each rater. The average rating was computed by dividing the raw score for each scale or index by the number of items in the respective scale or index. For example, the Initiate subscale on the parent version of the BRIEF is comprised of 8 items; whereas the Initiate subscale on the teacher version of the BRIEF is only comprised of 7 items. Thus, an average rating was necessary in order to compare the scores from the separate subscales. After deriving the average score for each parent- and teacher-completed subscale and index, paired sample *t* tests were computed to determine if there were significant differences between parent and teacher ratings. Effect sizes were calculated to compliment *p* values as a measure of parent and teacher rating differences, providing a metric for identifying substantive, versus simply statistical, significance. Effect sizes (d) at or above .20 are small, at or above .50 are medium, and above .80 are typically considered large (Cohen, 1988). Consequently, subscales or index scores with significant paired sample *t* values and large effect sizes represent parent and teacher disagreement, whereas nonsignificant paired sample *t* scores and small effect sizes suggest higher IRA. Table 2 summarizes these results. The mean differences between parents and teachers on the original BRIEF for the following subscale and index scores were statistically significant: Shift (d = 1.06), Emotional Control (d = .79), Plan/Organize (d = .65), Organization of Materials (d = 1.40), Behavior Regulation Index (d =

.95), Metacognitive Index (*d* = .69), and the Global Executive Composite (*d* = .85). The effect sizes for the original BRIEF ranged from 0.29 to 1.40 (mean = 0.74), with the majority (8 of 11 subscale/index) of effect sizes being in the medium to large range.

The mean differences between parent and teacher average ratings on the BRIEF-R resulted in significance for the Organization of Materials subscale (*d* = 1.04). Effect sizes ranged

Table 2

Differences Between Parent and Teacher Average Ratings Based on Scale Version

	Parent		Teacher			
Scale or Index	*M*	*(SD)*	*M*	*(SD)*	*t*	*d*
Original						
Inhibit	1.74	(0.41)	1.49	(0.59)	2.03	.50
Shift	1.78	(0.50)	1.36	(0.28)	3.85***	1.06
Emot Con	1.89	(0.48)	1.48	(0.53)	4.76***	.79
Initiate	1.66	(0.42)	1.45	(0.51)	1.52	.46
Work Mem	1.68	(0.46)	1.54	(0.54)	1.06	.29
Plan Org	1.62	(0.38)	1.39	(0.34)	2.50*	.65
Org Mater	2.13	(0.46)	1.44	(0.53)	4.11***	1.40
Monitor	1.79	(0.41)	1.59	(0.59)	1.73	.46
BRI	1.81	(0.37)	1.44	(0.40)	4.17***	.95
MI	1.74	(0.34)	1.48	(0.41)	2.46*	.69
GEC	1.77	(0.32)	1.47	(0.38)	3.20**	.85
Revised						
Inhibit	1.04	(0.67)	0.96	(1.21)	.33	.08
Shift	0.86	(0.69)	0.62	(0.63)	1.12	.37
Emot Con	1.10	(0.72)	0.70	(0.85)	2.02	.50
Initiate	1.04	(0.73)	1.04	(1.08)	.03	.00
Work Mem	1.04	(0.77)	1.01	(1.03)	.09	.03
Plan Org	0.99	(0.72)	0.83	(0.79)	.71	.21
Org Mater	1.87	(0.89)	0.85	(1.07)	2.86**	1.04
Monitor	1.34	(0.73)	1.13	(1.17)	.67	.21
BRI	1.01	(0.57)	0.76	(0.86)	1.22	.34
MI	1.19	(0.65)	0.98	(0.93)	.84	.27
GEC	1.12	(0.57)	0.89	(0.88)	.99	.31

Note: Emot Con = Emotional Control; Work Mem = Working Memory; Plan Org = Plan Organize; Org Mater = Organization of Materials; BRI = Behavior Regulation Index; MI = Metacognition Index; GEC = Global Executive Composite.
*p < .050, **p < .010, ***p < .001.

from 0.00 to a 1.04, with only the following two subscales that exceeded a small effect size score: Emotional Control (d = .50) and Organization of Material (d = 1.04). Moreover, the average effect size score for the revised BRIEF (d = .31) appeared to be lower than the average effect size score for the original BRIEF. These findings suggest that parents and teachers demonstrated greater disagreement (low IRA) on the original BRIEF compared to their ratings on the BRIEF-R. Although mean differences between parent and teacher ratings were less significant for the revised BRIEF, there appeared to be more variability in responding within each version (i.e., parent and teacher) of the revised BRIEF. That is, the standard deviations were larger for revised BRIEF. In particular, the teacher-completed revised BRIEF seemed to have the most variability of any other version with several subscale and index standard deviations exceeding +1 *SD* above the mean.

Correlations (IRR)

Although this study is primarily concerned with level of IRA between parents and teachers, IRR was also assessed to explore the total impact of the scale revision on "interrater similarity" (LeBreton & Senter, 2008). The present study examined IRR among parents and teachers using Pearson correlations between parents and teachers on the BRIEF and parents and teachers on the BRIEF-R. Correlations were also computed for parents on the BRIEF and BRIEF-R and teachers on the BRIEF and BRIEF-R. Due to the response format differences between the BRIEF and the BRIEF-R, Z-transformations were conducted. The Pearson correlations for the respective comparisons are displayed in Table 3. Correlations between the same raters' scores (e.g., parents) on the two rating scale versions (e.g., Parent BRIEF & Parent BRIEF-R) indicate a significant amount of consistency for both sets of raters. That is, both parents and teachers appeared to be ranking children in a similar order across the two versions.

Correlations between versions for parent ratings ranged from 0.41 to 0.84. Organization of Material was the only subscale with a nonsignificant correlation. Furthermore, both the Emotional Control and Monitor subscales yielded correlations at lower significance levels. Correlations for teacher ratings ranged from 0.60 to 0.87. The Shift (r = .60) subscale yielded lower significance values. A comparison of the average correlation for parents (r = .67) and average correlation for teachers (r = .78) suggests that the teachers' ratings were slightly more consistent across versions.

Whereas correlations between BRIEF versions among subscales completed by the same raters were primarily significant, correlations between raters for each version of the BRIEF were low and primarily nonsignificant. For the original BRIEF, parent-teacher correlations ranged from 0.10 to 0.74. Emotional Control (r = .74) and the Behavioral Regulation Index (r = .52)

Table 3

Pearson Correlations Between Parent and Teacher Ratings Using Multiple Versions

	Rating Scale Comparison			
	Parent	Teacher	Original	Revised
Scale or Index	Original – Revised	Original – Revised	Parent – Teacher	Parent – Teacher
Inhibit	0.84***	0.81***	0.43	0.48*
Shift	0.73***	0.60**	0.29	-0.06
Emot Con	0.55*	0.76***	0.74***	0.39
Initiate	0.74***	0.71***	0.10	0.06
Work Mem	0.75***	0.86***	0.30	0.04
Plan Org	0.69***	0.80***	0.34	0.14
Org Mater	0.41	0.76***	-0.18	-0.32
Monitor	0.64**	0.87***	0.31	-0.03
BRI	0.71***	0.76***	0.52*	0.24
MI	0.68***	0.85***	0.23	-0.05
GEC	0.67***	0.82***	0.32	-0.01

Note: Emot Con = Emotional Control; Work Mem = Working Memory; Plan Org = Plan Organize; Org Mater = Organization of Materials; BRI = Behavior Regulation Index; MI = Metacognition Index; GEC = Global Executive Composite.
*p < .050, **p < .010, ***p < .001.

were the only correlations that were statistically significant. Parent- teacher correlations on the BRIEF-R ranged from 0.01 to 0.48. The discrepancy between the average correlation on the BRIEF (r = .34) and the average correlation on the BRIEF-R (r = .16) indicates that parent and teacher ratings are more consistent with each other on the original version of the BRIEF. This finding does not indicate that agreement between parents and teachers is higher on the original version of the BRIEF. Rather, it suggests that, overall, the parent-teacher dyads rank-ordered the items for each child in a more consistent manner on the BRIEF compared to the BRIEF-R. It should be noted that while there appears to be greater consistency when rating children on the BRIEF, these correlations are modest at best.

Intraclass Correlations (ICC; IRR + IRA)

*ICC*s were estimated in the present study because they can account for absolute agreement (IRA) and rater consistency (IRR; LeBreton, Burgess, Kaiser, Atchley & James, 2003). *ICC*s will estimate both IRA and IRR when multiple subjects (e.g., children) are rated by a set of judges (e.g., parents and teachers). In this study, *ICC*s represent the proportion of observed variance in parent and teacher ratings that is due to the between-target (i.e., children) differences compared to the overall variance in ratings (LeBreton & Senter, 2008). Since an *ICC* is comprised of both IRA and IRR, high *ICC* values correspond to high agreement and relative consistency among parent and teacher ratings. Consequently, low *ICC* values may be a product of low IRA, IRR, or both (LeBreton et al., 2003). There are three separate cases to consider when a sample of children is being rated by a set of judges (e.g., parents and teachers). Consistent with Shrout and Fleiss (1979), a one-way analysis of variance (ANOVA) model was selected, because each child was rated by a different set of judges (case 1). A two-way ANOVA model did not seem viable in this study because k judges were not rating each target (case 2), and k judges were

not the only judges of interest (case 3). Furthermore, since analyses were not made at the item level, the average measures *ICC* were computed to account for multiple items and judges. Similar to the procedure implemented above, the average rating for each child is necessary in order to account for the different number of items that comprise each subscale or index. Thus, average ratings were utilized to calculate the *ICC*s. The *ICC*s for this study are displayed in Table 4. Overall, there seems to be a similar amount of IRA + IRR between the BRIEF and the BRIEF-R. However, a few of the individual *ICC*s for original BRIEF seems to be slightly higher than the respective *ICC*s for the revised BRIEF. For instance, the *ICC*s for the Working Memory and Monitor subscales, and the GEC index exceeded their revised counterpart (e.g., $r_{ICC} \geq .15$). This finding is not surprising given that *ICC*s account for IRR and the IRR for the BRIEF appeared to be higher than the IRR for the BRIEF-R.

Table 4

Intraclass Correlations (ICCs) Between Parent and Teacher Ratings

	Original	Revised
Scale or Index	Parent – Teacher	Parent – Teacher
Inhibit	.493	.593*
Shift	-.020	-.148
Emot Con	.667**	.484
Initiate	.128	.148
Work Mem	.417	.124
Plan Org	.369	.257
Org Mater	-.429	-.443
Monitor	.410	-.021
BRI	.370	.349
MI	.183	-.083
GEC	.190	-.005

Note: Emot Con = Emotional Control; Work Mem = Working Memory; Plan Org = Plan Organize; Org Mater = Organization of Materials; BRI = Behavior Regulation Index; MI = Metacognition Index; GEC = Global Executive Composite.
*p < .050, **p < .010, ***p < .001.

Version Preference

Both parents and teachers were asked to indicate their preference for one of the two versions (i.e., BRIEF vs. BRIEF-R). This information was obtained from a form asking parents and teachers to check a box next to the response format they preferred. An example of the form can be seen in Appendix B. The results indicated that 75% (n = 15) of the parents who participated in this study preferred the BRIEF-R, X^2 (1) = 5.0, $p < .05$, whereas only 55% (n = 11) of teachers preferred the BRIEF-R over the original version, X^2 (1) = 0.20, NS. Overall there seemed to be a slight preference for the revised version for both parent and teacher judges.

Discussion

The current study was designed to evaluate the impact of response format changes in parent and teacher rater agreement and reliability. Consistent with our hypothesis, the obtained results suggest that the change from the current response format to the frequency of behavior response format resulted in higher interrater agreement (IRA) on the revised version of the BRIEF. That is, mean differences between the average parent and teacher ratings on the original version of the BRIEF tended to be statistically significant; whereas fewer significant differences were obtained between average parent and teacher ratings on the revised version. Additionally, the majority of differences between parents and teachers on the original scale yielded medium to large effect sizes. In contrast, only two subscale differences exceeded a Cohen's *d* of .50 for average parent and teacher ratings on the revised BRIEF. Parents and teachers seem to agree more in their ratings of the target child's behavior when presented with a more precise metric on which to assess the frequency of behavior. Perhaps, the proposed frequency of behavior response format addresses and eliminates some of the ambiguity in rater interpretation of behavioral frequency. Sattler and Hoge (2006) corroborate this hypothesis with their suggestions for improving ratings recording.

Analyses of individual subscale and index discrepancies suggest that there are common significant differences across the two versions of the BRIEF. The discrepancies between the parent- and teacher-completed rating scales on the original version of the BRIEF yielded large effect sizes on both the *Emotional Control* (*d* = .79) and the *Organization of Materials* (*d* = 1.40)

subscales. Parent and teacher ratings on the revised version of the BRIEF produced a similar trend (d = .50 and d = .81, respectively). A potential explanation for these similarities is that task demands between the different environments (i.e., home vs. school) lead to very different behavioral presentations in the children being rated. Children may exhibit better emotional control and may be more organized at school versus at home, because the former potentially provides the child with more structure than the latter. The similarities in ratings across these two subscales might also have to do with the nature of the constructs they are attempting to assess; making it difficult for raters to assess these areas. The children in this sample may have had a real elevation of EF impairment across these two domains at home compared to school. This would also have to be assumed for all of the subscale and index scores. Another potential explanation for the similar findings across versions is that these two subscales represent a heterogeneous group of items. However, since the scope of this study was not concerned with the structural integrity of either version of the BRIEF this was not assessed further. Nonetheless, this still may remain a potential explanation as to why the average parent and teacher ratings on these two subscales resulted in low IRA on both versions (i.e., original and revised) of the BRIEF. Additional research focused on item-loadings and factor structure is needed to address this possibility. Interestingly, Baron (2000) noted the lower interrater agreement between parents and teachers on the same subscales (*Emotional Control* and *Organization of Materials*). The internal consistency of these two subscales appeared to be relatively high. Cronbach's alphas for the 10 *Emotional Control* and the 7 *Organization of Materials* items on the original BRIEF were .92 and .88 respectively; indicating that these scales are representing a unitary construct. This finding suggests that some factor(s) other than lack of internal consistency is likely accounting for the variations in ratings. Regardless of the explanation for the similar pattern of significant

differences across scale versions, there seems to be reasonable evidence that the IRA improved with the adaptation of the specific response format anchors.

While these findings should be considered preliminary, these data support the notion that response format on rating scales may influence IRA. Corroboration with subsequent research could have a significant impact on future scale development. In particular, developers of parent- and teacher-completed rating scales of EF should consider the revised format or a similar "frequency-based format" at the inception of rating scale development.

With respect to interrater reliability (IRR), the Pearson correlations reported above appear to be less conclusive in their support for the revised response format. While there tends to be strong relationship between versions completed by the same rater (e.g., teacher-completed original BRIEF and teacher-completed revised BRIEF), the same is not the case when the two different raters completed the same version (e.g., parent-completed original BRIEF and teacher-completed original BRIEF). This may not be a surprising finding considering that each rater is more likely to remain consistent with themselves since they are able to observe the child in the same environment. Regardless, the results did not support our prediction that the change in response format would lead to improved IRR. There appeared to be a slight advantage of the original version over the revised version of the BRIEF. Two of the scales on the original version of the BRIEF yielded significant correlations (*Emotional Control*, $r = .74$ and *BRI*, $r = .52$), whereas there was only one significant correlation when employing the revised response format (*Inhibit*, $r = .48$). Furthermore, there seems to be a trend of stronger correlations when comparing parent and teacher scores on the original version of the BRIEF than the revised version with the exception of the *Inhibit* subscale. In fact, the correlations between parents and teachers on the original BRIEF ($r = .43$) and the revised BRIEF ($r = .48$) for the *Inhibit* subscale were

remarkably consistent. Perhaps, difficulties associated with behavioral inhibition are both easier to recall for raters and consistent across both environments (i.e., home and school). The correlations between parent and teacher ratings on the original BRIEF (overall mean *r* = .34) were moderate and comparable to those reported in the BRIEF manual (overall mean r = .32; Gioia et al., 2000). Moreover, the authors reported that the correlation for the two subscales (*Initiate* and *Organization of Materials*) were notably lower, which is also the case with this sample (*r* = .10 and *r* = -.18 respectively). Gioia et al, (2000) suggest that these low correlations can be accounted for by the differences in the school and home environments; stating that teachers may aid students both in beginning tasks and organizing them.

These findings may indicate that the revised response format leads to lower IRR, or less consistent ratings, for parents and teachers. The revised format may provide parents and teachers with too many response options, making it difficult for raters to rank order the target child's behavior in a consistent manner. Another potential explanation for the minimal amount of IRR may again be related to the environmental differences faced by each rater. Some researchers (Mares et al., 2007) have suggested that teachers may be better equipped to discover EF deficits in children. This might be because teachers have the luxury of being able to compare the target child's behavior to the sample of children in their classrooms. The environmental difference may provide teachers with a more realistic perception of the developmental appropriateness of a behavior, whereas parents may not have access to a sample of children on which to base the frequency or appropriateness of their child's behavior. Thus, a teacher's rating may be more consistent because she has access to a normative comparison. Consequently, her original understanding of the response anchor *sometimes* may be more precise in relation to the parent's perception of that identical anchor. If this were the case, then it might explain why parent ratings

are less consistent across versions than teacher ratings. Furthermore, it may lend some insight as to why the correlations between parent and teacher ratings on the revised version are so weak, and in several cases negative, but still reach a higher level of IRA. Perhaps, once provided a common metric, parent and teacher ratings yield fewer mean differences, but are inconsistent due to the environmental task demands.

Since IRR for the revised version of the BRIEF was low, the *ICC*s for the revised BRIEF are more likely to be low as well. Similarly, there were low levels of IRA for parent and teacher raters on the original BRIEF subscales and indices. As discussed above, *ICC*s take into account both IRA and IRR. Thus, poor IRA, IRR, or both will likely result in weak *ICC*s because elevations in either within raters variance or between raters variance will decrease overall interrater reliability (LeBreton et al., 2003). Accordingly, the majority of *ICC*s between parent and teacher ratings were low and did not reach statistical significance for both the original and revised versions. The trend for higher *ICC*s on the original BRIEF most likely reflects the higher the overall stronger correlations reported above.

Although *ICC*s reflect both the level of absolute agreement and rater consistency, it is apparent that a high *ICC* can be acquired without the presence of both. Take, for example, the *Emotional Control* subscale on the original version. There was a relatively large difference (d = .79) between the average parent and teacher rating (i.e., low IRA), and a high level of rater consistency (r = .74), and the *ICC* still reached statistical significance despite the low IRA. Thus, low IRA is not sufficient by itself to result in a weak *ICC*. Potentially, this may be due to the fact that *ICC*s are essentially correlation coefficients themselves (Shrout & Fleiss, 1979) and may rely heavily on the consistency of the ratings in question. Being that the overall consistency of ratings was low to moderate (especially for the revised BRIEF), *ICC*s will likely be

nonsignificant. Further support for this discussion is highlighted by the fact that the two strongest *ICC*s on the revised BRIEF (i.e., *Inhibit* and *Emotional Control*) are also the two strongest pearson correlations. In general, there seems to be low interrater reliability, as measured by *ICC*s, across both the original and revised versions of the BRIEF; indicating a significant amount of within rater variance (lack of consistency) and between rater variance (lack of consensus). This finding implies that neither response format is resulting in a significant amount of interrater reliability. It should also be noted that *ICC*s are typically employed when a set of judges rate one or multiple targets (LeBreton et al., 2003; Shrout & Fleiss, 1979). It appears as if this study presents a unique scenario, in which two judges (e.g., parent and teacher) rate the same target (e.g., child), and the judges and target change with each rating.

Parents in this study reported a significant preference for the revised version of the BRIEF. In contrast, preference among teachers was equivocal. Since the data collection procedure for version preference only consisted of a forced-choice dichotomous selection between the original and revised scale versions, no formal conclusions regarding rater preference of each scale were formulated. Conceivably, teachers have access to normative comparisons in the classroom they find it easier to rate children using either version, whereas parents may have difficulty determining the appropriate response on the original version of the BRIEF given the ambiguity of the anchors used. Furthermore, the revised version provides parents with a specific behavior count or frequency, which may create the perception of more precise ratings. Future studies should systematically evaluate the reasons why some parents and teachers indicated a preference for the revised BRIEF over the original version. The fact that parents in this sample preferred the revised version of the BRIEF should be incentive to further explore the mechanisms driving parental preference. If future studies were able to replicate this finding it

could have an immediate impact on scale development as researchers and clinicians would most likely select a response format that was preferred by participants and patients respectively. Evaluating consumer satisfaction for the revised version of the BRIEF is a means to assess the social validity of the revised version of the BRIEF (Schwartz & Baer, 1991), and may ultimately lead to the development of rating scales with superior social validity.

There are several limitations of the study. Many of the sample characteristics limit the generalizability of our findings. The small sample size (n = 20) reduces the statistical power and increased the likelihood that the results may not be replicable. Power, or the likelihood of detecting a difference between conditions (e.g., rating scale versions), is impacted by sample size. Therefore there may be some real differences between the original and revised BRIEF that are not detected because of the small sample (Kazdin, 2003). Participants also volunteered or self-selected into the study and, as a result, were not selected randomly. The majority of the parent participants were Caucasian (75%), and perhaps most importantly, were recruited from the community. Since this study was not completed with a clinical sample, generalizations cannot be made to clinical populations. Consequently, the external validity of these findings must be taken into consideration. Results may have been different had children with documented EF impairment (e.g., ADHD, Autism, traumatic brain injury) been rated by two different raters. On the other hand, the increased severity of impairment may have led to higher rates of IRA, IRR, or both; but this is merely speculation and cannot be substantiated without further research.

Second, even though attempts were made to introduce more objectivity with the revised rating scale the nature of the data is still subjective. Parent- and teacher-completed rating scales require using a retrospective method of collecting data, which may be influenced by multiple characteristics. Although, there are clear limitations with any rating scale, there are certainly

advantages of rating scales over behavioral observations. Rating scales provide a vehicle for recording a wide set of behaviors and can be implemented for one or multiple individuals. Rating scales also assist in providing data that are more conducive for statistical analyses (Sattler & Hoge, 2006). However, there is no way to determine whether the revised response anchors actually increase the accuracy in ratings without having some type of direct observation (i.e., external validation) to accompany the parent and teacher ratings. Moreover, while attempts were made to have parents and teachers rate the target child at the same time, variability in the time that each rater completed the scales could have influenced ratings. Furthermore, we can't guarantee that parents and teachers used the time frame specified by the rating scale. Given that the main hypothesis of this study is that the scale revision will reduce ambiguity error, increase rater accuracy, and improve IRA, additional steps are needed to ensure whether the parent- and teacher-completed scales are actually more accurate, or if the revised scale is improving IRA at the expense of accurate ratings. Additional studies should attempt to address whether or not the frequency of behavior response format leads to more accurate ratings by complimenting the parent and teacher ratings with systematic observations of the target child's behavior or a meaningful outcome variable. Only then can firm conclusions be made about the improvement of rater accuracy.

Third, due to the nature of the items on the BRIEF two different types of frequency response formats were developed. Many of the items did not seem to fit into a discrete (i.e., behavioral count) response format so a continuous (i.e., percentage of time) response format was adopted. Although having two different types of anchors on the revised version of the BRIEF appeared necessary to capture the clinical characteristics of the item, it does add additional variance to ratings. Also, while we used a group validation process, assignment of items to

format may need to go through more rigorous evaluations. Consequently, the subscale and index scores are comprised of both continuous and discrete type response formats.

In summary, the revised version of the BRIEF appeared to increase IRA between raters, as measured by the infrequency of mean differences between parent and teacher average ratings. However, although the revised response format demonstrated superiority in IRA, there seemed to be a slight advantage of the original version over the revised version in IRR. That is, both versions demonstrated poor rater consistency, but the Pearson correlations tended to be larger for the original BRIEF. To reiterate, this study did not aim to suggest that separate raters should reach absolute agreement or perfect consistency. Rather, the aim was to examine the impact reducing the ambiguity error of the current response format of the BRIEF. Consequently, the frequency-based format utilized on the revised version of the BRIEF appears to deserve some merit for improving IRA. Additionally, the parents who participated in this sample indicated that they preferred the revised version over the original, lending support for the social validity of the revised response format.

The present study highlights the need for additional research related to scale development. In particular, future studies should focus on implementing this revised format with clinical populations and larger sample sizes. It is our hope that this study will serve as a catalyst for additional research pertaining to both rating scale development and assessment of clinical disorders associated with EF deficits. Future research must continue to manipulate and evaluate measures in order to develop a rating scale that accurately taps into the construct(s) of interest as well increases the objectivity of overall ratings.

References

Achenbach, T. M. (1991). *Integrative Guide to the 1991 CBCL/4-18, YSR, and TRF Profiles*. Burlington, VT: University of Vermont, Department of Psychology.

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(2),* 213-232.

Aiken, L. R. & Groth-Marnat, G. (2006). *Psychological testing and assessment* (12th ed.). New York: Allyn & Bacon.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.) Washington, DC: Author.

Anastasi, A. & Urbina, S. (1997) *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Bagwell, C. L., Molina, B. S. G., Pelham, W. E. Jr., & Hoza, B. (2001). Attention-deficit hyperactivity disorder and problems in peer relations: Predictions from childhood to adolescence. *Journal of the American Academy of Child & Adolescent Psychiatry, 40(11),* 1285-1292.

Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. *Psychological Bulletin, 121,* 65-94.

Barkley, R. A., Fischer, M., Smallish, L., & Fletcher, K. (2002). The persistence of attention-deficit/hyperactivity disorder into young adulthood as a function of reporting source and definition of the disorder. *Journal of Abnormal Psychology, 111,* 109-121.

Barkley, R. A. (2006). *Attention-deficit hyperactivity disorder: A handbook for diagnosis and treatment* (3rd ed.). New York: Guilford Press.

Baron, I. S. (2000). Test review: Behavior rating inventory of executive function. *Child Neuropsychology, 6(3),* 235-238.

Biederman, J., Faraone, S. V., Milberger, S., Curtis, S., Chen, L., Marrs, A, et al. (1996). Predictors of persistence and remission of ADHD into adolescence: Results form a four-year prospective follow-up study. *Journal of American Academy of Child and Adolescent Psychiatry, 35(3),* 343-351.

Biederman, J., Mick, E., Faraone, S. V., Braaten, E., Doyle, A., Spencer, T., et al. (2002). Influence of gender on attention deficit hyperactivity disorder in children referred to a psychiatric clinic. *American Journal of Psychiatry, 159,* 36-42.

Blake, D. D., Weathers, F. W., Nagy, L. M., Kaloupek, F. D., et al. (1995). The development of a clinician-administered PTSD scale. *Journal of Traumatic Stress, 8(1),* 75-89.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64(4),* 410-421.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Conners, C. K. (1997). *Conners' Rating Scales Revised.* North Tonawanda, NY: Multi-Health Systems.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan Executive Function System (DKEFS): Examiner's manual*. San Antonio, TX: The Psychological Corporation.

Donders, J. (2002). The Behavior Rating Inventory of Executive Function: Introduction. *Child Neuropsychology, 8(4),* 229-230.

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). *The Behavioral Rating Inventory of Executive Function professional manual*. Odessa, FL: Psychological Assessment Resources.

Hinshaw, S. P. (1987). On the distinction between attentional deficits/hyperactivity and conduct problems/aggression in child psychopathology. *Psychological Bulletin, 101,* 443–463.

Holland, M. L., Gimpel, G. A., & Merrell, K. W. (1998). Innovations in assessing ADHD: Development, psychometric properties, and factor structure of the ADHD symptoms rating scale (ADHD-SRS). *Journal of Psychopathology and Behavioral Assessment, 20(4),* 307-332.

Holland, M. L., Gimpel, G. A., & Merrell, K. W. (2001) *ADHD Symptoms Rating Scale.* Wilmington, Delaware: Wide Range, Inc.

Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed.). Boston, MA: Allyn & Bacon.

Kenny, D. A. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review, 98,* 155-163.

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K. P., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 8(1),* 80-128.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11(4),* 815-852.

Lockwood, K. A., Marcotte, A. C., & Stern, (2001). Differentiation of attention-deficit/hyperactivity disorder subtypes: Application of a neuropsychological model of attention. *Journal of Clinical and Experimental Neuropsychology, 23*, 317–330.

Mahone, E. M., Cirino, P. T., Cutting, L. E., Cerrone, P. M., et al. (2002). Validity of the Behavior Rating Inventory of Executive Function in children with ADHD and/or Tourette Syndrome. *Archives of Clinical Neuropsychology, 17*, 643-662.

Mares, D., McLuckie, A., Schwartz, M. & Saini, M. (2007). Executive function impairments in children with Attention-Deficit Hyperactivity Disorder: Do they differ between school and home environments? *The Canadian Journal of Psychiatry, 52(8),* 527-534.

McCandless, S. & O'Laughlin, L. (2007). The clinical utility of the Behavior Rating Inventory of Executive Function (BRIEF) in the diagnosis of ADHD. *Journal of Attention Disorders, 10(4),* 381-389.

McMahon, R. J. & Frick, P. J. (2007). Conduct and Oppositional Disorders. In E. J. Mash & R. A. Barkley (Eds.), *Assessment of Childhood Disorders* (4th ed., pp. 132-183). New York: Guilford Press.

Merrell, K. W. (2008). *Behavioral, social, and emotional assessment of children and adolescents* (3rd ed.). New York: Lawrence Erlbaum Associates.

Molina, B. S. G. & Pelham, W. E. Jr. (2003). Childhood predictors of adolescent substance use in a longitudinal study of children with ADHD. *Journal of Abnormal Psychology, 112(3),* 497-507.

Reid, R. & Maag, J. W. (1994). How many fidgets in a pretty much: a critique of behavior rating scales for identifying students with ADHD. *Journal of School Psychology, 32(4),* 339-354.

Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavioral Assessment System for Children manual.* Circle Pines, MN: American Guidance Service.

Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior Assessment System for Children* (2nd ed.). Circle Pines, MN: American Guidance Services.

Sattler, J. M., & Hoge, R. D. (2006). *Assessment of children: Behavioral, social, and clinical foundations* (5th ed.). San Diego, CA: Jerome M. Sattler Publisher Inc.

Schwartz, I. S., & Baer, D. M. (1991). Social validity assessments: Is current practice state of the art? *Journal of Applied Behavioral Analysis, 24(2),* 189-204.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86(2),* 420-428.

Smith, B. H., Barkley, R. A., & Shapiro, C. J. (2006). Attention-Deficit/Hyperactivity Disorder. In E. J. Mash & R. A. Barkley (Eds.), *Treatment of Childhood Disorders* (3rd ed., pp. 65-136). New York: Guilford Press.

Smith, B. H., Barkley, R. A., & Shapiro, C. J. (2007). Attention-Deficit/Hyperactivity Disorder. In E. J. Mash & R. A. Barkley (Eds.), *Assessment of Childhood Disorders* (4th ed., pp. 53-131). New York: Guilford Press.

Sullivan, J. R. & Riccio, C. A. (2007). Diagnostic group differences in parent and teacher ratings on the BRIEF and Conner's scales. *Journal of Attention Disorders, 11(3),* 398-406.

Tinsley, H. E. A. & Weiss, D. J. (2000). Interrater reliability and agreement. In H. E. A. Tinsley & S. D. Brown (Eds.). *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95-124). San Diego: Academic Press.

Weathers, F. W., Newman, E., Blake, D. D., Nagy, L. M., et al. (2004). *Interviewer's Guide to the Clinician-Administered PTSD Scale.* Los Angeles, CA: Western Psychological Services.

Welsh, M. C. & Pennington, B. F. (1988). Assessing frontal lobe functioning in children: Views from developmental psychology. *Developmental Neuropsychology, 4,* 199-230.

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V. & Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: A meta-analytic review. *Biological Psychiatry, 57(11),* 1336-1346.

Appendices

Appendix A

Demographic questionnaire.

Participant Contact Information/Demographics Questionnaire

Notice: This information will remain strictly confidential and will be destroyed at the completion of the study.

Parent Name: _____

Child Name: _____

Parent Home Phone: _____

Parent Cell Phone: _____

Parent Work Phone: _____

Parent Email: _____

Address: _____

_____

I prefer to be contacted by _____ Phone _____ Email _____ Either

Teacher's Name: _____

School Name: _____

Teacher's Email: _____

Would you be interested in being contacted for future studies conducted only by Dr. Shapiro or his graduate students? Your contact information would be retained in a secured cabinet until you wish to remove yourself. Your answer will not influence your eligibility to participate in this research study.

_____ Y _____ N

Participant Contact Information/Demographics Questionnaire

Notice: This Information will remain strictly confidential
and will be destroyed at the completion of the study

Code: _____

Child Age: _____

Child Date of Birth

(DD/MM/YYYY) ____ / ____ / ____

Child Sex: (circle one)

Male Female

Child Race: (circle one)

African American	Native American
Asian	Mixed (specify) _____
Caucasian	Other (specify) _____
Hispanic	

Is your child taking any prescription medicine for psychological/psychiatric issues?

Yes No

Have there been any changes in his/her medication within the last 3 months? (circle one)

Yes No

Has your child changed schools or teachers within the last 3 months? (circle one)

Yes No

Please indicate whether the family has experienced any of the following stressors in the past 3 months. (circle one)

Marital:	Yes	No
Health:	Yes	No
Relocation:	Yes	No
Employment/Financial:	Yes	No

If you have concerns about the impact these stressors have had on your family, a referral list is attached for your convenience.

Appendix B

Preference form.

Thank you for choosing to participate in our study. In an attempt to improve rating scales we would like know which of the two rating scales you preferred. Please check the box next to the format that you preferred.

☐ **N = Never**
S = Sometimes
O = Often

OR

☐ **0 = never**
1 = once or twice a week
2 = three to four times a week
3 = almost every day or daily
4 = almost every hour or hourly

0 = none of the time
1 = very little of the time (less than 10%)
2 = some of the time (≈20% - 30%)
3 = much of the time (≈50% - 60%)
4 = most of the time (more than 80%)

Appendix C

Table 5. Nonparametric Summary Table

	Spearman Rank Correlations				Wilcoxon Matched Pairs	
	Parent	Teacher	Original	Revised	Z scores	
	Original–Revised	Original–Revised	Parent–Teacher	Parent–Teacher	Original	Revised
Scale or Index						
Inhibit	.92***	.72***	.51*	.47*	-1.88	-.95
Shift	.77***	.68***	.31	-.20	-3.08**	-1.21
Emo Cont	.54**	.74***	.83***	.45*	-3.27**	-2.02*
Initiate	.70***	.71***	.18	-.01	-1.65	-.30
Work Mem	.75***	.83***	.20	-.09	-1.22	-.36
Plan Org	.64**	.79***	.37	-.01	-3.32*	-.71
Org Mat	.38	.83***	-.24	-.32	-3.17**	-2.54*
Monitor	.52*	.85***	.39	-.03	-1.96	-.99
BRI	.66**	.71***	.77***	.18	-3.02**	-1.49
MI	.61**	.88***	.38	-.09	-2.33*	-1.14
GEC	.52*	.79***	.52*	-.15	-2.73*	-1.31

Note: Emot Con = Emotional Control; Work Mem = Working Memory; Plan Org = Plan Organize; Org Mater = Organization of Materials; BRI = Behavior Regulation Index; MI = Metacognition Index; GEC = Global Executive Composite.
*p < .050, **p < .010, ***p < .001.