Published in final edited form as:

Child Neuropsychol. 2020 January ; 26(1): 113–128. doi:10.1080/09297049.2019.1617262.

Parent versus Teacher Ratings on the BRIEF-Preschool Version in Children with and without ADHD

Heather Schneider¹, Matthew Ryan¹, E. Mark Mahone^{1,2}

¹Kennedy Krieger Institute, Baltimore, MD;

²Johns Hopkins University School of Medicine, Baltimore, MD

Abstract

Caregiver rating scales represent an important component of comprehensive child neuropsychological assessments for conditions such as Attention-deficit/Hyperactivity Disorder (ADHD); however, low inter-rater reliability (parent vs. teacher) often complicates interpretation. It has been challenging to identify the factors contributing to inter-rater variability, particularly when parents and teachers complete slightly different versions of the same rating scale. The present study examined the associations between parent- and teacher-reported executive functions in 84 children, ages 4–5 years, with and without symptoms of ADHD, using the Behavior Rating Inventory of Executive Function-Preschool (BRIEF-P). Use of the BRIEF-P allows for direct comparison of symptom ratings because parents and teachers complete the exact same measure. Significant associations between raters were observed on 4 of 5 BRIEF-P subscales when rating children with ADHD, but on only 1 subscale when rating typically developing (TD) children. The Shift scale in particular displayed low, non-significant inter-rater association in both groups. Significant group-by-rater interactions were observed for Working Memory and Plan/Organize scales, and driven by larger inter-rater *T*-score discrepancies in the TD group, such that teachers rated children as having more symptoms than parents. Conversely, examination of raw scores reflected no significant rater differences in the TD group, but significant or nearly significant differences on multiple scales in the ADHD group, such that parents rated more symptoms than teachers. Inter-rater associations for the BRIEF-P appear to vary based on who is being rated (i.e., children with or without ADHD), the specific subscales, and whether standardized or raw scores are analyzed.

Keywords

Executive Function; Childhood; Neuropsychology; Assessment; Validity

There are a variety of assessment methods available to clinicians when assessing an individual for executive function weaknesses and/or Attention-deficit/Hyperactivity Disorder (ADHD). Among neuropsychologists, executive functions (EF) are typically assessed through four primary methods: individual performance-based tests of EF, inter-test contrast scores, intra-test “process” variables, and caregiver rating scales (Mahone & Slomine, 2007).

Each approach has its own conceptual strengths, as well as empirically identified weaknesses. Unfortunately, performance-based tests, inter-test contrast scores, and process variables show limited correlations with “real-life” or “day-to-day” behavioral differences in functioning (Beebe, Ris, & Dietrich, 2000; Bodnar, Prahme, Cutting, Denckla, & Mahone, 2007; Ezpeleta, Granero, Penelo, de la Osa, & Domènech, 2015; Mahone, Cirino, et al., 2002; Wodka et al., 2008)

It has been argued that the fourth method, rating scales, have the most ecological validity because they are a report of daily functioning in a natural setting, and by people who interact with the individual on a daily basis. Multiple raters (e.g., self, parent, teacher) provide multiple perspectives on the child’s functioning. The ecological validity of rating scales are not impacted in the same ways as performance measures by factors such as optimal testing environment and use of compensatory strategies that may allow an individual to perform better on performance measures than they are able to perform in their typical environment (Chaytor & Schmitter-Edgecombe, 2003). Rating scales have been useful in identifying real-life executive dysfunction in the absence of impaired performance on EF tasks in the clinical setting (Davidson, Cherry, & Corkum, 2016; Mahone, Cirino, et al., 2002; Mahone & Hoffman, 2007; Mahone, Zabel, Levey, Verda, & Kinsman, 2002). In addition to being more reflective of daily executive weaknesses, rating scales of EF have shown stronger associations with frontal lobe volumes, compared to associations between frontal volumes and performance-based measures (Mahone, Martin, Kates, Hay, & Horská, 2009).

Regardless of the primary method of assessment that a clinician utilizes, assessment in multiple settings is important in determining level of impairment, and is required in the clinical diagnosis of disorders, such as ADHD. The practicality of clinicians’ direct observation of individuals across settings is limited; thus, rating scales are frequently employed to obtain parent, teacher, or other caregiver impressions of the individual in multiple real-life settings. Unfortunately, assessment can become complicated when multiple raters provide different observations. Across multiple populations, parent and teacher ratings of attention, behavioral, and executive functioning typically display low to moderate correlations (Achenbach, Ivanova, & Rescorla, 2017; Chevignard, Soo, Galvin, Catroppa, & Eren, 2012; Lavigne, Dahl, Gouze, LeBailly, & Hopkins, 2015; Lavigne, Gouze, Hopkins, & Bryant, 2016; Mares, McLuckie, Schwartz, & Saini, 2007; Moens, Weeland, Van der Giessen, Chhangur, & Overbeek, 2018; Murray et al., 2007; Narad et al., 2015; van der Ende, Verhulst, & Tiemeier, 2012; Wochos, Semerjian, & Walsh, 2014; Wolraich et al., 2004). These studies suggest that low inter-rater reliability can be a result of a number of factors, such as informant pair, age of child, gender of child, inherent biases of the rater, level of parental stress, negative family experiences, and behavior type being rated. Raters in similar roles/contexts tend to correlate higher than those in different roles (Achenbach, Ivanova, & Rescorla, 2017) and informant pairs involving a teacher as one of the raters tend to correlate less than parent and self-report pairs (van der Ende, Verhulst, & Tiemeier, 2012).

The low rates of inter-rater reliability, particularly between parents and teachers, can be problematic for assessment of younger children who are typically less able to recognize and describe their own behavioral difficulties, and thus reliant on raters who may be influenced by a number of factors outside of the explicit presenting behaviors. More specifically, in

young children (sample age ranges 4–6 and 4–8 years), rater discrepancies were related to factors within the parent-child relationships, such as parenting stress related to child temperament, parent-child dysfunctional interactions, and negative parenting behaviors (Chen et al., 2017; Moens et al., 2018). In addition, boys' behavior tended to be rated worse by parents than by observers, while girls' behavior tended to be rated better by parents than observers (Moens et al., 2018).

Awareness and consideration of rater discrepancies and the relevant contributing factors is important in the diagnostic process and subsequent accessibility to care. In a longitudinal study of children ages 3–14, higher inter-rater reliability was shown to better indicate the trajectory of externalized behavior problems over time (Roskam, 2017). Alternatively, lower inter-rater reliability in children grades K-4 can lead to decreased diagnostic rates that result from enforcement of the criterion of multiple settings of impact (Wolraich et al., 2004), potentially missing a young child who may have benefited from early interventions. Because it has the potential to impact early intervention and treatment accessibility, it is important to establish an understanding of inter-rater reliability in scales developed for the assessment of young children, not just in regard to the presence and/or absence of diagnostic symptoms, but also in regard to symptoms and behaviors associated with disorders frequently diagnosed in childhood. EF skills are one of these associated areas that are necessary in nearly all aspects of an individual's daily functioning. Investigating the properties of EF rating scales is of particular importance because not only are EF difficulties associated with common disorders identified in childhood, but this is an area in which it has been previously established that performance-based assessment does not necessarily identify real-life EF difficulties resulting in increased reliance on rating scales.

An additional factor complicating inter-rater reliability discussed less frequently in the literature is that rating scales are often designed specifically for the rater and context, with different item sets and normative data. While utilizing different item sets can be a benefit, in that the scales are designed to assess traits and behaviors specifically relevant to the setting in which the individual is observed, it also prevents the clinician from determining if there truly are differences in the way the raters are viewing the individual, or if the differences are a result of the items themselves.

The Behavior Rating Inventory of Executive Function-Preschool (BRIEF-P; Gioia, Espy, & Isquith, 2003) is a widely used rating scale that can be completed by parents and teachers to aid in the identification of EF difficulties in children ages 3–5 years. The BRIEF-P is of particular interest in studying parent-teacher reliability because it is designed to be utilized by multiple raters to assess preschool age children, while allowing the raters to respond to exactly the same items, rather than setting-specific items. This consistency allows investigation of inter-rater reliability between high discrepancy informant pairs (i.e., parent-teacher), within an age range at which reliable self-report is typically unattainable, while controlling for item set differences.

Investigations of the measurement properties of the BRIEF-P have demonstrated acceptable internal consistency and convergent validity but inconsistent findings regarding replication of the developers' factor structure in non-referred and clinical samples (Bonillo, Araujo

Jiménez, Jané Ballabriga, Capdevila, & Riera, 2012; Duku & Vaillancourt, 2014; Isquith, Gioia, & Espy, 2004; Skogan et al., 2016; Spiegel, Lonigan, & Phillips, 2017). Unfortunately, the majority of research on the BRIEF-P is conducted using parent ratings, resulting in little information regarding how multiple informant ratings compare when rating the same children. The limited research involving parent and teacher report on the BRIEF-P is consistent with the research on other rating scales and research on the BRIEF (childhood version), in that low to modest parent-teacher agreement is observed (Chevignard et al., 2012; Gioia et al., 2003; Mares, McLuckie, Schwartz, & Saini, 2007). In addition, differences in parent and teacher ratings of the same children have resulted in an expanded dimensional model, but one in which not all of the same items cluster together across raters (Duku & Vaillancourt, 2014). In regard to the efficacy of the BRIEF-P in clinical evaluation, there is a high degree of association between parent responses on the BRIEF-P and ADHD symptoms (McGoey et al., 2000; Mahone & Hoffman, 2007; Skogan et al., 2015).

While rating scales are useful tools in the assessment of EF difficulties, particularly in identifying “real-world” weaknesses, there is still a great deal of variability between raters, and very little is known about the use of rating scales to assess EF in the preschool age range. Even less is known about how the BRIEF-P functions when comparing parent and teacher report. Investigation of the BRIEF-P can provide valuable information regarding multiple informant ratings of behavior because unlike other rating scales, use of the BRIEF-P allows for direct comparison of symptoms in multiple settings by allowing parents and teachers to complete the exact same measure. The current study is thus unique because it directly compares parent and teacher reports of behavior in both a clinical (ADHD) and a typically developing sample. It is important to investigate and compare the measurement properties of rating scales in both clinical and typically developing groups to determine if the measure functions similarly between groups, as unforeseen differences could inadvertently impact accessibility to treatment. Many of the above mentioned factors influential on inter-rater reliability (e.g., parenting stress, parent-child relationships) are potentially different in a clinical population than a typically developing population as a result of the child’s behavioral difficulties. EF difficulties are commonly observed in children with ADHD, making investigation of EF within this population particularly relevant. In addition, because a high degree of correlation between parent ratings of EF and ADHD symptomatology has been previously demonstrated, research would suggest a strong presence of EF difficulties in the clinical group under investigation, limiting potential floor effects. Based upon that research one would also expect limited variability among parent ratings within the ADHD group based on symptomatology of the disorder itself, allowing us to more specifically investigate variability between raters within that group. We hypothesized that parent and teacher ratings would diverge within and across groups. We also hypothesized that there would be greater parent/teacher concordance in a clinical group because of the sensitivity of BRIEF-P to symptoms seen in ADHD.

METHODS

Study Procedures

The Johns Hopkins Medicine Institution Review Board granted approval for the study. Participants were recruited from advertisements in the community, pediatricians' offices, and local daycare centers, to participate in a longitudinal study of brain and behavioral development in young children. After description of the study, parents of participants signed written informed consent, and child participants provided verbal assent. Participants were initially screened via telephone interview with a parent to determine eligibility. Once enrolled, participants completed a neuropsychological assessment battery that included the cognitive, language, and inhibitory control measures. Parents (and teachers, if available) also completed rating scales assessing behavior and executive function at the time of testing. The focus of the present study was on rating scale measures collected at initial study visit.

Participants

Inclusion and Exclusion Procedures.—The exclusion criteria for this study were the same as used for the longitudinal study. Children enrolled in the larger longitudinal study were required to be between the ages of 4 years, 0 months through 5 years, 11 months at study enrollment. Participants were excluded if they had any of the following, established via review of medical/developmental history, and/or by study screening assessment: 1) diagnosis of Intellectual Disability or Autism Spectrum Disorder; 2) known visual impairment; 3) treatment of any psychiatric disorder (other than ADHD) with psychotropic medications [for those with diagnosis of ADHD, treatment with stimulants was allowed, whereas children treated with other psychotropic medications were excluded]; 4) any history of DSM-IV or DSM-5 Axis I diagnosis other than Oppositional Defiant Disorder or Adjustment Disorder; 5) neurological disorder (e.g., epilepsy, traumatic brain injury, tic disorder); 6) documented hearing loss ≥ 25 decibels loss in either ear; 7) reported history of physical, sexual, or emotional abuse; 8) Full Scale IQ scores < 80 (as determined by previous assessment or study screening assessment). In addition, children were excluded if there was a history of a Developmental Language Disorder (DLD) either determined during the initial phone screen, based on prior assessment (completed within one year of the current assessment), or determined during screening visit. DLD exclusion was made in deference to evidence that language impairments may influence development of inhibitory control, response preparation, and working memory—core elements of executive function often affected in ADHD (Hagberg, Miniscalco, & Gillberg, 2010).

Diagnostic methods for the ADHD and typically developing (TD) groups were adapted from the NIH Preschoolers with Attention-Deficit/Hyperactivity Disorder Treatment Study (PATS; Kollins et al., 2006; Posner, 2007). For 4-year olds, diagnosis of ADHD was made using modified DSM-IV-TR criteria, based on parent report on the Diagnostic Interview Schedule for Children-Young Child (YC-DISC; Lucas, 1998; Lucas, Fisher, & Luby, 2008) or Diagnostic Interview for Children and Adolescents, Fourth Edition—DICA-IV (Reich, 1997), depending on age, and the DSM-IV ADHD Scales (Scales L: Inattention and M: Hyperactive-Impulsive) of the Conners' Parent Rating Scales-Revised (CPRS-R; Conners, 1997). The YC-DISC is a highly structured, computer-assisted diagnostic instrument that

assesses common psychiatric disorders, as defined by DSM-IV, that present in young children. The DICA-IV is the parallel version of the computer-assisted, structured interview for older children and adolescents. In the present study, the YC-DISC was used for 4-year olds and the DICA-IV was used for 5-year olds. To be included in the ADHD group, symptoms must have been present for at least 6 months, and cross-situational impairment (defined as parent report of problems at home and with peers—as not all children were enrolled in school) was required. Additionally, children in the ADHD group were required to have *T*-scores ≥ 65 on one or both of the DSM-IV ADHD Scales (Scales L and M) of the CPRS-R or the Conners' Teacher Rating Scales-Revised (CTRS-R; Conners, 1997).

Once children met general entry/exclusion criteria above, they were included in the control group only if they did not meet categorical diagnostic criteria for ADHD on the YC-DISC or DICA-IV. Additionally, children in the TD group were required to have *T*-scores < 65 on the CPRS-R and CTRS-R DSM-IV ADHD Scales.

Having met the general inclusion/exclusion criteria, participants were selected for the current study based on the criterion that they have both parent and teacher report on the BRIEF-P at the time of their initial study visit. Children who did not have responses from both raters were excluded from the current analyses, but were not excluded from the overall longitudinal study.

Study Measures

Behavior Rating Inventory of Executive Function- Preschool Version (BRIEF-P; (Gioia et al., 2003)).—The BRIEF-P is a rating scale for children ages 2–5 years designed to be completed by parents/caregivers or teachers that assesses executive behaviors in daily environments. Raters rate the child's behavior on a three-point Likert scale (Never=0, Sometimes=1, Often=2). Parent and teacher forms are identical, allowing for direct comparison in the present study.

The BRIEF-P includes 63 items and is organized into five clinical scales (Inhibit, Shift, Emotional Control, Working Memory, Plan/Organize), three clinical indexes (Inhibitory Self-Control, Flexibility, and Emergent Metacognition), and a Global Executive Composite. Raw score totals are converted to age-referenced *T*-scores. The five clinical scale scores served as the dependent variables of interest for the present study. Higher *T*-scores indicate caregiver ratings suggesting greater difficulty in that area of executive functioning. Internal consistency and reliability for the five scales is good ($\alpha > .80$ for all scales for parent ratings and $\alpha > .90$ for all scales for teacher ratings). In contrast, published agreement between parent and teacher BRIEF-P ratings is weaker (mean agreement = .19).

Conners' Parent- and Teacher-Rating Scales-Revised-Long Form (CPRS-R; CTRS-R; Conners, 1997).—Dimensional ratings of ADHD symptom severity were obtained using the DSM-IV oriented scales from the CPRS-R and the CTRS-R, including Scale L (DSM-IV Inattentive) and Scale M (DSM-IV Hyperactive/Impulsive). These scales were used to determine eligibility status for the two clinical groups. Internal consistency for ages 3–5 years on these scales using Chronbach's alpha ranged from .86 to .94.

Clinical Evaluation of Language Functions-Preschool-2 (CELF-P; Wiig, Secord, & Semel, 2004).—The CELF-P is an individually administered, norm-referenced test developed to identify language and communication disorders in preschool children. Participants scoring < -1.5 SD on either the Receptive Language or Expressive Language Index of the CELF-P, or < -1.0 SD on *both* indices, were excluded. The CELF-P Core Language Index was used to characterize the sample demographics.

Wechsler Preschool and Primary Scale of Intelligence-Third Edition (WPPSI-III; Wechsler, 2002).—The WPPSI-III is a widely used measure of early cognitive abilities. The Full Scale IQ score (FSIQ) was used in determining participant eligibility and to characterize the sample.

Hollingshead Index (Hollingshead, 1975).—Socioeconomic status (SES) for each participant was estimated by a widely used four-factor index (i.e., gender, marital status, education, and occupation).

Data Analysis Plan

Demographic variables were compared using one-way ANOVAs (for dimensional variables) and chi-square analyses (for categorical variables) to assess differences between clinical groups. Associations between parent and teacher ratings were examined separately within ADHD and TD groups using paired samples correlations. Group (ADHD vs. TD) and rater (parent vs. teacher) differences, as well as group-by-rater interaction effects, were examined using repeated measures ANOVAs for each BRIEF-P scale (*T*-score mean). Paired samples *t*-tests were subsequently conducted for each Index and raw score total to determine differences between raters within each group. Raw scores were included in analyses because they provide additional information that cannot be obtained through standardized scores alone. Raw scores can be used to track individual changes in longitudinal research and clinical interventions by examining the raw level of item endorsement. Although the BRIEF-P items are identical across raters, different normative data sets are used, therefore having the potential to influence inter-rater reliability. Because of the different normative symptoms and the potential to utilize raw scores to track individual change, it is important to investigate the relationship of raw scores between raters and groups when scales are identical across raters. Significance level for group comparisons was set at $p=.01$ to control for multiple comparisons across each of the five subtests within each analysis.

RESULTS

Demographics

Sample demographics are listed in Table 1. The sample included 84 children, ages 4–5 years ($M = 4.99$, $SD = .57$), including 35 typically developing children (18 boys) and 49 children with clinical symptoms of ADHD (29 boys). None of the participants were taking stimulant medication, and all were medication naïve at the time of participation. The sample was predominantly Caucasian (90%), with an additional 5% of the sample identifying as African-American. The remaining 5% identified as Asian (3%), multiple races (1%), and Other (1%). There were no significant group differences in the distribution of race

($\chi^2_{(2)}=4.70, p=.32$), sex ($\chi^2_{(2)}=.498, p=.31$), age, SES, core language skills, or FSIQ (Table 1). The sample used in the study ($n=84$) included only children who had both full parent and teacher BRIEF-P forms completed. An additional 43 children were recruited for the larger study but not included in the present analyses because they did not have both parent and teacher forms (often because the child was not yet in a formal preschool setting). There were no significant differences in age ($p=.441$), SES ($p=.096$), or FSIQ ($p=.470$) between children included in the present sample and those excluded for lack of both parent and teacher BRIEF-P forms.

BRIEF-P Scale Scores: Inter-rater Associations—Paired samples correlations between parent and teacher report for each scale (T -score means) were conducted separately for ADHD and TD groups, with results listed in Table 2. These analyses revealed that within the ADHD group, there were significant associations between parent and teacher ratings on four of the five scales (correlations ranging from .30 to .34), with only the Shift scale showing non-significant inter-rater association ($r=-.01$). In contrast, within the TD group, only the Plan/Organize scale showed significant inter-rater associations. Using Fisher's r -to- z transformation to examine differences in magnitude of inter-rater associations between ADHD and TD groups, none of the differences in associations between groups reached statistical significance (Table 2).

Group-by-Rater Interactions—Results of repeated measures ANOVAs comparing group and raters for each of the five scales (T -score means) are summarized in Table 3. There were no statistically significant effects of rater for any of the scale score totals. As would be anticipated, there were significant effects of group on all five scales, with scores significantly higher (showing more impairment) in the ADHD group ($p<.001$ for all). Analysis of group-by-rater interaction effects revealed significant interactions for two scales: Working Memory, and Plan/Organize (Table 3). Of note, the effect size for group differences (ADHD vs. TD) for these two scales was essentially twice the magnitude for parent ratings ($\eta_p^2=.591$, and .505 respectively) compared to teacher ratings on the same scales ($\eta_p^2=.250$, and .201 respectively).

Following the finding of significant group-by-rater interactions on two scales, direct contrasts of ratings (parent vs. teacher) on each scale were examined using paired samples t -tests (Table 4). Both group-by-rater interactions were driven by stronger effects (i.e., larger inter-rater discrepancies) in the TD group, compared to the ADHD group. Teacher rating on the Inhibit scale was also significantly higher than parent rating in the TD group. Of note, within the TD group, teacher scores were consistently higher than parent ratings on each corresponding scale, with an average overall difference of 4.0 T -score points.

Recognizing that parent/teacher differences in standardized score totals may reflect differences in norms, an additional set of inter-rater comparisons (paired samples t -tests) was made using *raw score* totals (Table 5). Since parent and teacher forms of the BRIEF-P are identical, these contrasts allow for direct comparison of actual number of items endorsed. Interestingly, examining raw scores revealed a slightly different pattern of inter-rater discrepancies than examining T -scores. Using raw scores, there were no significant inter-rater differences on any scale within the TD group. Conversely, within the ADHD group,

significant inter-rater differences were observed on the Plan/Organize scale; in this case, *parents* rated more problems than teachers on the same scale. Specifically, across all scales, within the ADHD group, parents rated their children higher (i.e., more executive dysfunction) than did the child's teacher by an average of 2.3 raw score points.

DISCUSSION

The BRIEF-P is a unique behavioral rating scale because it uses identical forms for both parent and teacher responses, thus allowing for direct inter-rater comparisons. The current findings, however, suggest that relationships between parent and teacher ratings on the BRIEF-P scales are more complex than the inter-rater correlations outlined in the technical manual (Gioia et al., 2003), and depend on factors such as who is being rated, the scale itself, and whether raw or standardized scores are analyzed. For the BRIEF-P, associations between rater responses appear to vary depending on whether the child being rated has symptoms of ADHD, or is more typically developing. In our analyses, significant inter-rater correlations were observed on the majority of BRIEF-P scales when rating children with ADHD, whereas fewer significant associations were identified when rating typically developing children, possibly because of the reduced range of observed scores among the typically developing group. At the same time, the pattern of inter-rater associations varied as a function of the type of score analyzed (raw vs. *T*-score) as the overall pattern using raw scores was reversed, reflecting greater differences between raters of children with ADHD.

Across all scales, the *Shift* scale was most discrepant between raters, showing the lowest inter-rater reliability, with correlations never reaching significance in either group. The inconsistency among raters for the Shift scale may be because this construct is less well characterized overall, or possibly because it is less directly applicable to preschoolers (compared to Inhibit or Working Memory). Because many of the behaviors assessed on the Shift scale are so new to preschool children in social settings, there may be an inherent adult expectation that some of the items, such as adjusting to new people, taking time to become comfortable, hesitating to join new activities, and feeling overstimulated in large crowds are more expected to occur at this age. If these behaviors are anticipated, they are viewed as less problematic in comparison to working memory or inhibitory weaknesses that are apparent across both novel and familiar settings, and likely interfering more with daily activities and behavioral expectations. Setting specific differences may also contribute to inconsistency between raters. After a period of adjustment, the school and teacher become familiar to the majority of students and routines are established. The teacher is then observing responses to established routines in a single, familiar setting, whereas parents are observing responses in a variety of novel settings, in addition to the familiar routines of home. Significant inconsistencies in environment and the type of "shift" demands required of the child may contribute to inconsistencies between raters as well.

When comparing group differences for scale means, parents and teachers rated the same children discordantly on the Working Memory and Plan/Organize scales; though, the direction and magnitude of those discrepancies depended on whether the children were typically developing or displayed symptoms of ADHD. In fact, when considering *standardized* scores, the group-by-rater interactions were driven primarily by differences in

rater responses of typically developing children, such that teacher ratings were, on average, approximately 6 *T*-score points higher than parent ratings on each of the two corresponding scales, as well as on the Inhibit scale. Nevertheless, it should be noted that scale score means for typically developing children were consistently within normal limits, showing little variability, and less than 20% had any instance of elevated scale ratings ($T > 65$ or above) by either rater.

Conversely, among preschool children with symptoms of ADHD, parent/teacher rater differences in standardized score scale means were small, with both rater groups consistently identifying executive dysfunction across most scales (again, with the exception of the Shift scale, which was relatively insensitive to identifying dysfunction in the current sample). The exception to these rater similarities for children with ADHD was on the Working Memory scale, which approached significance and for which parents rated their children an average of 6 *T*-score points higher than did teachers of the same children. Higher parent ratings on this scale may be due to differences in setting-specific expectations, and what is viewed as problematic. Preschool teachers interact with multiple children of variable abilities, and in a potentially more distracting setting. As a result, it is reasonable to assume that repetition, redirection, and other similar behaviors that support working memory requirements are a natural part of introducing young children to novel learning experiences in a group setting. In contrast, parents may be asking the child to independently complete familiar tasks (e.g., activities of daily living (ADLs)) in a less distracting setting, and expecting the child to complete these tasks with fewer reminders and redirection. There may also be fewer points of reference (e.g., other children) in the home setting for children to get themselves back on task should they forget; thus, the reminders become necessary. In this context, working memory difficulties might be more noticeable and problematic in the home setting.

For standardized scores, the observed patterns for parent/teacher discordance were clearly driven by the different norm sets used to calculate *T*-scores in this age range (4–5 years). At the low end of the score distribution for each scale (*T*-scores < 55), similar raw score totals produced *higher* *T*-scores for teacher norms, compared to parent norms, suggesting that in this age and problem range, teacher reported data in the standardization sample may be more “conservative” with regard to reporting EF concerns. Conversely, at the higher end of the distribution (*T*-scores > 65), parent and teacher norms appeared to converge such that similar raw score ratings led to largely equivalent *T*-score means. This observation does not suggest, however, that the magnitude of parent and teacher scale totals (raw scores) were similar. As with standard scores, parents did *not* rate children similarly to teachers when considering raw scores. This observation was particularly true among those with ADHD for whom ratings by parents identified greater executive dysfunction than teachers (showing clinical relevance), although short of reaching a conservative level of statistical significance. These results suggest that at the most basic level, while both parents and teachers agree that preschool children with ADHD display significant problems with executive function, parents are rating these children as displaying a greater relative degree of executive difficulties than teachers are rating these same children.

There are a number of possible explanations that could be speculated regarding the higher parent ratings among children with ADHD. Of note, at the time of assessment, none of the

young children in the present sample with symptoms of ADHD had ever taken stimulant medication, which may have contributed to exacerbation of some of their behaviors. Additionally, parents of children in this age range may spend more time with their children in the home setting during the course of a typical day, resulting in increased opportunity to witness behavioral impairment. Conversely, preschool or Kindergarten teachers may have structured individual or class-based behavioral plans, classroom strategies, and/or routines already in place that assist young students with executive control, perhaps minimizing the degree of apparent functional impairment in children with ADHD. Because direct instruction of skills related to executive control functions (e.g., following rules, waiting one's turn) is commonly part of the preschool curriculum, most students (ADHD and typically developing) are just beginning to learn these skills and how they applied in larger group settings. During this process, the learning curve for young students may not make executive function weaknesses as immediately apparent to teachers, whereas at home, parents may be experiencing increased frustration with a young child with ADHD may display disruptive behaviors that impede the accomplishment of routine ADL tasks (e.g., dressing, mealtimes, hygiene activities).

The absence of significant rater differences in actual raw scores among typically developing children may have been a function of measurement issues specific to the instrument, leading to a restricted distribution. In general, the BRIEF-P is designed to capture abnormality, rather than strengths in executive functioning. Although there were differences in *T*-scores on some scales (all within normal limits), parents and teachers were not rating TD children differently at the most basic level because their executive functioning was not seen as problematic. In other words, a child with "average" executive functioning and a child with "exceptional" executive functioning would be rated the same (i.e., "0") on scales designed to capture problematic behavior because neither child is displaying a problem, resulting in a potential floor effect. Thus, while it is possible to obtain a *T*-score as high as 106 (i.e., 5.6 standard deviations above the mean), it is only possible to obtain a *T*-score as low as 34 (i.e., only 1.6 standard deviations below the mean).

Overall, the results of this study suggest that when comparing raters on dimensional measures, such as the BRIEF-P, examination of overall associations and mean differences addresses different questions. Significant correlations are found between rater responses; however, the significance and the magnitude of the correlation may depend on other clinical variables, such as group membership and behavioral scale. As such, a strength of the present study was the use of the BRIEF-P to investigate these patterns of relationships. Because parents and teachers complete the exact same rating form rather than separate versions for each rater, as is the case with most other behavioral rating measures (including all other versions of the BRIEF rating scales), our design allowed comparison of both raw and standard scores. This allowed investigation of potential differences between each scale's raw total, controlling for the different standardization samples for each rater. A limitation of the current study that is difficult to address is the potential floor effect when examining ratings of typically developing children because most behavioral rating scales of this type are designed to capture weaknesses or abnormality, rather than strengths. In addition, although this study controlled for a number of demographic factors related to the child, it did not control for factors specific to the raters such as gender of rater, the duration/quality of

relationships between child and rater, rater stress, or inherent biases, which have been shown to impact ratings (Achnebach, Ivanova, & Rescorla, 2017).

Future research using the BRIEF-P should seek to analyze the relationships between parent and teacher responses at the item level. For example, item level analysis would also provide a closer examination of the Shift scale to help identify which aspects of that scale result in the observed inconsistencies between raters. Limitations that could be addressed in the future include an increased sample size and inclusion of children in the ADHD group with comorbid diagnoses to better generalize to the wider population of children with ADHD. In addition, a larger sample size, perhaps with a larger age range (in particular with more younger children), would allow for better comparisons at each age group to see if inter-rater reliability changes throughout this period of development. While this study did not investigate more specific rater demographics and the quality of rater-child relationships, as has been reflected in previous research, future research using the BRIEF-P would further findings in those areas by removing the confound of separate rating scales.

In conclusion, this study expanded on previous research regarding inter-rater reliability in the assessment of young children. The majority of previous research has focused on aspects of the person, either the rater or the rated, or the relationships between some combination of these people. The findings of those studies are confounded by the fact that the rating scales for each rater are different. The current study investigated the use of a single measure in a rating pair that has been shown to be particularly problematic in regard to inter-rater reliability, parents and teachers (van der Ende, Verhulst, & Tiemeier, 2012), but yet still a pair frequently used in assessment of children. These results support that when utilizing a rating scale in the assessment of EF in young children, the diagnostic information obtained varies based on the raters and those being rated, but also suggests variability related to aspects of the measure itself (e.g., the defined scales, standardized scores vs. raw scores). While the scales correlate, they act differently in different groups, and may actually be measuring different things. These results are important clinically and from a research perspective. Clinically, it is important to understand that factors distinct from child behavior may impact inter-rater reliability, which subsequently impacts diagnosis, intervention access, and behavioral trajectory (Roskam, 2017; Wolraich et al., 2004). This is particularly important in an area of functioning that is difficult to directly assess through performance measures and relies on rating scales to assess real-life EF difficulties. (Davidson, Cherry, & Corkum, 2016; Mahone, Cirino, et al., 2002; Mahone & Hoffman, 2007; Mahone, Zabel, Levey, Verda, & Kinsman, 2002). From a research perspective, data obtained can also be influenced by the heterogeneity of the population under study, and how the data are analyzed (raw vs. standardized scores) which can impact the pattern of findings and interpretation of results if these differences are not considered. Clinicians and researchers should understand how rater and measure effects might moderate their conclusions in identifying executive functioning difficulties and differences in children with and without ADHD.

ACKNOWLEDGEMENTS

A portion of this study was presented at the Annual Meeting of the International Neuropsychological Society in New Orleans, LA, February 2, 2017. Supported by R01 HD068425, U54 HD079123, UL1 RR025005, the Johns Hopkins Brain Sciences Institute, and the Kennedy Krieger Institute Women's Initiative Network.

References

- Achenbach TM, Ivanova MY, & Rescorla LA (2017). Empirically based assessment and taxonomy of psychopathology for ages 1½–90+ years: Developmental, multi-informant, and multicultural findings. *Comprehensive Psychiatry*, 79, 4–18. doi:10.1016/j.comppsy.2017.03.006 [PubMed: 28356192]
- Beebe DW, Ris MD, & Dietrich KN (2000). The relationship between CVLT-C process scores and measures of executive functioning: lack of support among community-dwelling adolescents. *Journal of Clinical and Experimental Neuropsychology*, 22, 779–792. doi:10.1076/jcen.22.6.779.950 [PubMed: 11320436]
- Bodnar LE, Prahme MC, Cutting LE, Denckla MB, & Mahone EM (2007). Construct validity of parent ratings of inhibitory control. *Child Neuropsychology*, 13, 345–362. doi: 10.1080/09297040600899867 [PubMed: 17564851]
- Bonillo A, Araujo Jiménez EA, Jané Ballabriga MC, Capdevila C, & Riera R (2012). Validation of Catalan version of BRIEF-P. *Child Neuropsychology*, 18, 347–355. doi: 10.1080/09297049.2011.613808 [PubMed: 21961875]
- Chaytor N, & Schmitter-Edgecombe M (2003). The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychology Review*, 13, 181–197. doi: 10.1023/B:NERV.0000009483.91468.fb [PubMed: 15000225]
- Chen Y-C, Hwang-Gu S-L, Ni H-C, Liang SH-Y, Lin H-Y, Lin C-F, ... Gau SS-F (2017). Relationship between parenting stress and informant discrepancies on symptoms of ADHD/ODD and internalizing behaviors in preschool children. *PLOS One*, 12. doi:10.1371/journal.pone.0183467
- Chevignard MP, Soo C, Galvin J, Catroppa C, & Eren S (2012). Ecological assessment of cognitive functions in children with acquired brain injury: A systematic review. *Brain Injury*, 26, 1033–1057. doi:10.3109/02699052.2012.666366 [PubMed: 22715895]
- Conners CK (1997). *Conners' Rating Scales - Revised Technical Manual*. North Tonawanda, New York: Multi-Health Systems Inc.
- Davidson F, Cherry K, & Corkum P (2016). Validating the Behavior Rating Inventory of Executive Functioning for children with ADHD and their typically developing peers. *Applied Neuropsychology: Child*, 5, 127–137. doi:10.1080/21622965.2015.1021957 [PubMed: 26110907]
- Duku E, & Vaillancourt T (2014). Validation of the BRIEF-P in a sample of Canadian preschool children. *Child Neuropsychology*, 20, 358–371. doi:10.1080/09297049.2013.796919 [PubMed: 23721091]
- Ezpeleta L, Granero R, Penelo E, de la Osa N, & Domènech JM (2015). Behavior Rating Inventory of Executive Functioning-Preschool (BRIEF-P) applied to teachers: Psychometric properties and usefulness for disruptive disorders in 3-year-old preschoolers. *Journal of Attention Disorders*, 19, 476–488. doi:10.1177/1087054712466439 [PubMed: 23264366]
- Gioia GA, Espy KA, & Isquith PK (2003). *The Behavior Rating Inventory of Executive Function-Preschool Version (BRIEF-P)*. Odessa, FL: Psychological Assessment Resources.
- Hagberg BS, Miniscalco C, & Gillberg C (2010). Clinic attenders with autism or attention-deficit/hyperactivity disorder: Cognitive profile at school age and its relationship to preschool indicators of language delay. *Research in Developmental Disabilities*, 31, 1–8. [PubMed: 19713073]
- Hollingshead A (1975). *Four factor index of social status*. New Haven, CT: Yale University, Department of Sociology.
- Isquith PK, Gioia GA, & Espy KA (2004). Executive function in preschool children: examination through everyday behavior. *Developmental Neuropsychology*, 26, 403–422. doi:10.1207/s15326942dn2601_3 [PubMed: 15276902]
- Kollins S, Greenhill L, Swanson J, Wigal S, Abikoff H, McCracken J, ... Bauzo A (2006). Rationale, design, and methods of the Preschool ADHD Treatment Study (PATs). *Journal of the American Academy of Child and Adolescent Psychiatry*, 45, 1275–1283. [PubMed: 17023869]
- Lavigne JV, Dahl KP, Gouze KR, LeBailly SA, & Hopkins J (2015). Multi-domain predictors of oppositional defiant disorder symptoms in preschool children: Cross-informant differences. *Child Psychiatry & Human Development*, 46, 308–319. doi:10.1007/s10578-014-0472-4 [PubMed: 24997089]

- Lavigne JV, Gouze KR, Hopkins J, & Bryant FB (2016). Multi-domain predictors of attention deficit/hyperactivity disorder symptoms in preschool children: Cross-informant differences. *Child Psychiatry & Human Development*, 47, 841–856. doi:10.1007/s10578-015-0616-1 [PubMed: 26669698]
- Lucas CP, Fisher P., Luby JL (1998). Young Child DISC-IV Research Draft: Diagnostic Interview Schedule for Children. New York, NY: Columbia University, Division of Children Psychiatry, Joy and William Ruane Center to Identify and Treat Mood Disorders.
- Lucas CP, Fisher P, & Luby JL (2008). Young Child DISC-IV: Diagnostic Interview Schedule for Children. New York, NY: Columbia University, Division of Children Psychiatry, Joy and William Ruane Center to Identify and Treat Mood Disorders.
- Mahone EM, Cirino PT, Cutting LE, Cerrone PM, Hagelthorn KM, Hiemenz JR, ... Denckla MB (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. [PubMed: 14591848]
- Mahone EM, & Hoffman J (2007). Behavior ratings of executive function among preschoolers with ADHD. *The Clinical Neuropsychologist*, 21, 569–586. [PubMed: 17613979]
- Mahone EM, Martin R, Kates WR, Hay T, & Horská A (2009). Neuroimaging correlates of parent ratings of working memory in typically developing children. *Journal of the International Neuropsychological Society*, 15, 31–41. doi:10.1017/S1355617708090164 [PubMed: 19128526]
- Mahone EM, & Slomine BS (2007). Managing dysexecutive disorders In Hunter S & Donders J (Eds.), *Pediatric Neuropsychological Intervention* (pp. 287–313). Cambridge, UK: Cambridge University Press.
- Mahone EM, Zabel TA, Levey E, Verda M, & Kinsman S (2002). Parent and self-report ratings of executive function in adolescents with myelomeningocele and hydrocephalus. *Child Neuropsychology*, 8(4), 258–270. [PubMed: 12759823]
- 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? *Canadian Journal of Psychiatry*, 52, 527–534. [PubMed: 17955916]
- McGoey KE, Bradley-Klug K, Crone D, Shelton TL & Radcliffe J (2000). Normative data of the ADHD-Rating Scale IV-Preschool version. Paper presented at the National Association of School Psychologists annual convention, New Orleans, LA.
- Moens MA, Weeland J, Van der Giessen D, Chhangur RR, & Overbeek G (2018). In the eye of the beholder? Parent-observer discrepancies in parenting and child disruptive behavior assessments. *Journal of Abnormal Child Psychology*. doi:10.1007/s10802-017-0381-7
- Murray DW, Kollins SH, Hardy KK, Abikoff HB, Swanson JM, Cunningham C, ... Chuang SZ (2007). Parent versus teacher ratings of attention-deficit/hyperactivity disorder symptoms in preschoolers with attention-deficit/hyperactivity disorder treatment study (PATS). *Journal of Child and Adolescent Psychopharmacology*, 17, 605–619. doi:10.1089/cap.2007.0060 [PubMed: 17979581]
- Narad ME, Garner AA, Peugh JL, Tamm L, Antonini TN, Kingery KM, ... Epstein JN (2015). Parent-teacher agreement on ADHD symptoms across development. *Psychological Assessment*, 27, 239–248. doi:10.1037/a0037864 [PubMed: 25222436]
- Posner K, Melvin GA .. Murray DW .. Gugga SS .. Fisher P .. Skrobala A .. Cunningham C .. Vitiello B .. Abikoff HB .. Ghuman JK .. Kollins S .. Wigal SB .. Wigal T .. McCracken JT .. McGough JJ .. Kastelic E .. Boorady R .. Davies M .. Chuang S (2007). Clinical presentation of attention-deficit/hyperactivity disorder in preschool children: the preschoolers with attention-deficit/hyperactivity treatment study (PATS). *Journal of Child and Adolescent Psychopharmacology*, 17, 547–562. [PubMed: 17979577]
- Reich W, Welner Z .. & Herjanic B (1997). The Diagnostic Interview for Children and Adolescents-IV. North Tonawanda: Multi-Health Systems.
- Roskam I (2017). The clinical significance of informant agreement in externalizing behavior from Age 3 to 14. *Child Psychiatry & Human Development*. doi:10.1007/s10578-017-0775-3
- Skogan AH, Egeland J, Zeiner P, Øvergaard KR, Oerbeck B, Reichborn-Kjennerud T, & Aase H (2016). Factor structure of the Behavior Rating Inventory of Executive Functions (BRIEF-P) at age

three years. *Child Neuropsychology*, 22, 472–492. doi:10.1080/09297049.2014.992401 [PubMed: 25573465]

Skogan AH, Zeiner P, Egeland J, Urnes A-G, Reichborn-Kjennerud T, & Aase H (2015). Parent ratings of executive function in young preschool children with symptoms of attention-deficit/hyperactivity disorder. *Behavioral and Brain Functions*, 11. doi:10.1186/s12993-015-0060-1

Spiegel JA, Lonigan CJ, & Phillips BM (2017). Factor structure and utility of the Behavior Rating Inventory of Executive Function—Preschool Version. *Psychological Assessment*, 29, 172–185. doi:10.1037/pas0000324 [PubMed: 27148785]

van der Ende J, Verhulst FC, & Tiemeier H (2012). Agreement of informants on emotional and behavioral problems from childhood to adulthood. *Psychological Assessment*, 24, 293–300. doi: 10.1037/a0025500 [PubMed: 21928910]

Wochos GC, Semerjian CH, & Walsh KS (2014). Differences in parent and teacher rating of everyday executive function in pediatric brain tumor survivors. *The Clinical Neuropsychologist*, 28, 1243–1257. doi:10.1080/13854046.2014.971875 [PubMed: 25343533]

Wodka EL, Mostofsky SH, Prahme C, Gidley Larson JC, Loftis C, Denckla MB, & Mark Mahone E (2008). Process examination of executive function in ADHD: Sex and subtype effects. *The Clinical Neuropsychologist*, 22, 826–841. [PubMed: 18609314]

Wolraich ML, Lambert EW, Bickman L, Simmons T, Doffing MA, & Worley KA (2004). Assessing the impact of parent and teacher agreement on diagnosing attention-deficit hyperactivity disorder: *Journal of Developmental & Behavioral Pediatrics*, 25, 41–47. doi: 10.1097/00004703-200402000-00007 [PubMed: 14767355]

Table 1.

Participant Demographic Comparisons

	ADHD (<i>n</i> =49)		TD (<i>n</i> =35)			
	Mean	SD	Mean	SD	<i>p</i>	η_p^2
Age	5.0	0.6	4.9	0.5	.450	.007
SES	56.7	10.3	59.1	9.0	.286	.015
FSIQ	108.4	11.6	109.7	13.2	.612	.003
CELF-P	106.3	9.2	105.5	11.8	.747	.001

Note: SES = Hollingshead Index total; FSIQ = Wechsler Preschool and Primary Scale of Intelligence, Third Edition (WPPSI-III) Full Scale IQ; CELF-P = Clinical Evaluation of Language Functions-Preschool Version, Core Language Index

Table 2.

Inter-rater Pearson Correlations Between Parent and Teacher BRIEF-P Ratings

		ADHD Group	TD Group	ADHD vs. TD
	BRIEF-P Scale	<i>r</i>	<i>r</i>	<i>p</i> (one-tailed)
T-Scores	Inhibit	.337*	.222	.295
	Shift	-.013	-.073	.397
	Emotional Control	.299*	.179	.291
	Working Memory	.318*	.104	.164
	Plan/Organize	.314*	.376*	.378

* *Note:* $p < .05$ (two-tailed)

Table 3. Rater and Group Differences in Scale *T*-Scores (Repeated Measures ANOVA)

Scale	Main Effect Rater		Group-by-Rater Interaction	
	<i>p</i>	η_p^2	<i>p</i>	η_p^2
Inhibit	.127	.028	.021	.064
Shift	.587	.004	.599	.003
Emotional Control	.300	.013	.240	.017
Working Memory	.919	.001	.001	.135
Plan/Organize	.317	.012	.005	.094

Note: Group differences (TD < ADHD) were significant ($p < .001$) for all scales

Table 4. Parent vs. Teacher Ratings by Group (*T*-Score Means)

ADHD Group (<i>n</i> = 49)						
	Parent		Teacher			
Scale	Mean	SD	Mean	SD	<i>p</i>	<i>d</i>
Inhibit	72.6	11.3	71.2	16.9	.590	0.2
Shift	56.6	11.6	54.9	11.7	.474	0.2
Emotional Control	63.7	18.3	59.2	14.9	.122	0.5
Working Memory	72.7	14.4	66.7	16.4	.025	0.7
Plan/Organize	68.1	14.4	64.8	17.0	.214	0.4
TD Group (<i>n</i> = 35)						
Inhibit	44.4	8.1	50.7	9.4	.002	1.2
Shift	45.0	7.9	45.0	6.5	.987	0.0
Emotional Control	47.8	12.2	48.1	11.9	.913	0.1
Working Memory	44.1	6.9	50.6	9.2	.002	1.2
Plan/Organize	43.3	8.4	50.1	10.4	.001	1.3

Table 5.

Parent vs. Teacher Ratings by Group (Raw Score Totals)

ADHD Group (n = 49)						
	Parent		Teacher			
Scale	Mean	SD	Mean	SD	p	d
Inhibit	36.9	6.8	35.2	10.2	.246	0.3
Shift	17.0	4.7	15.2	4.4	.053	0.6
Emotional Control	19.6	6.0	17.4	6.2	.029	0.7
Working Memory	34.6	7.3	31.8	9.8	.050	0.6
Plan/Organize	20.9	4.4	18.0	5.7	.001	1.0
TD Group (n = 35)						
Inhibit	20.5	4.7	22.4	6.8	.113	0.6
Shift	12.4	3.1	11.5	2.5	.215	0.4
Emotional Control	14.0	4.0	12.8	5.0	.228	0.4
Working Memory	20.0	3.4	21.9	6.1	.118	0.6
Plan/Organize	12.8	2.6	13.1	3.7	.614	0.2