7-19-2017

Generalizability of Teacher Ratings of Attention-Deficit/ Hyperactivity Disorder Symptoms Across Rater and Form

Meredith Sanders Manguno

GENERALIZABILITY OF TEACHER RATINGS OF ATTENTION-DEFICIT/HYPERACTIVITY DISORDER SYMPTOMS ACROSS RATER AND FORM

by

Meredith Sanders Manguno

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Major: Psychology

The University of Memphis

August 2017

Abstract

This study investigated the dependability of teacher ratings of Attention-Deficit/Hyperactivity Disorder characteristics through the completion of the Attention-Deficit/Hyperactivity Disorder Rating Scale-IV, School Version and the Attention-Deficit/Hyperactivity Disorder Rating Scale-5, School Version, Child Form. Teacher pairs from five classrooms across two schools each independently rated the same ten randomly selected students from their classroom roster and completed three ratings scales on each student, one of which was intended as a filler. Generalizability theory was used to examine error variance across instrument, rater, and classroom concurrently, resulting in strong dependability coefficients for Inattention composite, Hyperactive/Impulsive composite, and Total Score composites. Variance estimates showed the differences in how teachers approached the completion of the ratings scales on a particular student and accounted for the largest proportion of total variance across all three composites. The results of assessing multiple sources of error concurrently instead of in isolation, as well as the outcomes of the recently updated ADHD Rating Scale-5 in this study, show promise in using Generalizability Theory to drive decisions for teachers and instrument users.

Table of Contents

Generalizability of Teacher Ratings of Attention-Deficit/Hyperactivity Disorder Symptoms Across Rater and Form

Attention-deficit/hyperactivity disorder (ADHD) is the most common childhood disorder and is characterized by "persistent, impairing, and developmentally inappropriate behaviors of inattention, hyperactivity, and impulsivity" (American Psychiatric Association [APA], 2013, p. 61). ADHD is associated with greater risks for "low academic achievement, poor school performance, school suspensions and expulsions, poor peer and family relations, aggression, conduct problems, and delinquency" (Barkley, 2014, p. 169). It is a well-researched topic, and the preponderance of the literature indicates that the way this chronic condition manifests itself often varies quite a bit due to individual differences of the person. Often, children are not diagnosed until they begin attending school or until school becomes challenging. It is common to see a direct correlation between intensity of symptoms and levels of impairment in childhood (Sibley & Pelham, 2011), usually reported by parents and teachers of these children.

According to the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5;* APA, 2013), to be diagnosed with ADHD, a child must exhibit at least six symptoms related to inattention, six symptoms related to hyperactivity–impulsivity, or a combination of both inattention and hyperactivity-impulsivity symptoms. Inattention symptoms include failing to pay close attention to details or making careless mistakes, appearing distracted when being directly spoken to, having difficulty with organization of tasks and activities, avoiding tasks that require sustained mental effort, and being forgetful during daily activities. Hyperactivity–impulsivity symptoms include fidgeting or difficulty remaining still, engaging in inappropriate or excessive movement, acting as if "driven by a motor" or being unable to engage in self-control, talking excessively, and experiencing difficulty waiting. Any symptoms related

to ADHD are required to have been persistent in nature (lasting for at least six months) and to have interfered with the child or adolescent's development or ability to function across two or more settings (e.g., at home and at school). The DSM-5 (APA, 2013) also requires several of the symptoms to have been present prior to age 12 and to not be part of another disorder. If enough inattention and the hyperactivity–impulsivity symptoms are present, the child or adolescent can be diagnosed with the combined type presentation of ADHD.

Review of Literature

Multiple methods can be used during the assessment process to evaluate a child or adolescent for ADHD characteristics.

Identifying ADHD Using Behavior Rating Scales

The best way to assess whether children and adolescents meet the diagnostic criteria for ADHD is through the collaborative use of multiple sources of information (APA, 2013; Barkley, 2006), and behavior rating scales completed by parents and teachers are particularly useful during this assessment process (Pelham, Fabiano, & Massetti, 2005). Behavior rating scales provide a consistent format for developing conclusions on the behavioral characteristics of a child (Whitcomb & Merrell, 2013). Their advantages include providing data on low-frequency behaviors, requiring fewer costs (in terms of training and time) than other assessment techniques, yielding more reliable data than unstructured interviews, acquiring results based on a longer period of time in a natural environment, and involving knowledgeable informants about the child or adolescent (Whitcomb & Merrell, 2013). Many parent and teacher rating scales are used to target ADHD symptoms, including the Parent and Teacher forms of the Behavioral Assessment System for Children, Third Edition (BASC-3; Reynolds & Kamphaus, 2015); the Child Behavior Checklist and Teacher Report Forms from the Achenbach System of Empirically Based

Assessment (ASEBA; Achenbach & Rescorla, 2001); Conners-3 Parent and Teacher Report Forms (Conners, 2008); Parent and Teacher Forms of the Vanderbilt Assessment Scale (Wolraich, 2002); and the School and Home Versions of the ADHD Rating Scale-5 (DuPaul, Power, Anastopoulos, & Reid, 2016).

Teacher ratings targeting these behaviors are particularly useful in providing an objective method of assessing the problem in a short amount of time. In addition, low frequency or rare problem behaviors can be assessed through behavior ratings scales in a way that observations or interviews would not accurately evaluate. Further, the ratings provide key information across natural environments, such as the student's classroom, while using the judgments and knowledge of an expert informant, such as the student's teacher.

Challenges in Employing Behavior Rating Scales with Teachers

A referral to special education is most likely to occur when the child enters school, as the classroom brings new requirements with typical behavior expectations that are needed for children to learn at school (Sax & Kautz, 2003). Norfolk and Floyd (2016) further explained how teachers are in a unique situation, as they are more likely to identify attention or hyperactivity/impulsivity problems due to the behavioral expectations of the child.

Pelham et al. (2005) stated that the classroom teacher is often the initial individual to identify and report complaints about ADHD. This statement by Pelham et al. has been supported by studies of children with ADHD and ratings in classroom settings that document a greater display of problematic characteristics typical of students with ADHD. This finding further highlights the need for the evaluation of academic functioning and classroom behavior in a comprehensive assessment, especially for ADHD.

Despite these advantages, behavior rating scales are not without limitations. Whitcomb and Merrell (2013) described the measurement problems related to rating scales as being between two classes, which are bias of response and error variance. Bias of response refers to "the way informants completing the rating scales potentially may create additional error by the way they use the scales" (p. 130). However, the more commonly discussed class is error variance, which often overlaps with bias of response in problems of rating scale measurement, and will be discussed further below.

Volpe, Briesch, and Gadow (2011) reported sources of error as including rater, item, time, method, setting, and dimension. However, there are well-known confounds in the use of behavior rating scales, which include the previously mentioned error variance. More specifically, Whitcomb and Merrell (2013) described the types of error variance attributable to the use of behavior rating scales, which include instrument, source, and setting variance.

Instrument Variance

If there are two different rating scales, both targeting inattention or hyperactivity, that are not producing similar scores, this is most likely attributable to instrument variance. Bergeron, Floyd, McCormack, and Farmer (2008) described instrument variance as "inconsistencies between scores that supposedly assess the same construct yielded by different rating scales administered concurrently" (p. 94). This source of variance could be related to response format, such as scoring items on frequency rather than severity, or item content, which could be interpreted differently with the use of words based on an outdated diagnostic manual (Whitcomb & Merrell, 2013).

There is no standard for confirming concurrent validity, which involves obtaining and comparing scores and information around the same time (AERA, APA, & NCME, 2014), but the

level of agreement should be moderate or higher (correlation coefficients ranging from .50 to .80; De Los Reyes & Kazdin, 2005; DuPaul, Power, McGoey, Ikeda, & Anastapoulos, 1998; Kolko & Kazdin, 1993). For example, the Conners Teacher Rating Form Inattention and Hyperactivity/Impulsivity subtests (Conners et al., 1998) and the ADHD Rating Scale-IV School Version (DuPaul et al., 1998) correlated at .80 to .88. In addition, the mean correlation between the ASEBA Teacher Rating Form (Achenbach & Rescorla, 2001) and the Conners-3 teacher scale (Conners, 2008) was reported as .80 (Pelham et al., 2010), which is above moderate.

Source and setting variance

The variance attributable to different sources and settings has been seen throughout literature over the last two decades. Teachers often have discrepancies in their ratings on a scale due to response bias, which can be problematic when these discrepancies are used in the evaluation of an assessment of a child (De Los Reyes & Kazdin, 2004). Studies have investigated these discrepancies between informants' ratings by examining how the same child was rated by two or more informants on evaluations of child behaviors (Chi & Hinshaw, 2002; De Los Reyes & Kazdin, 2004; Youngstrom et al., 2000). De Los Reyes and Kazdin (2004) found that differences could be influenced by the characteristics of the samples, as well as the measure of discrepancy, which was determined through a systematic review focusing on ratings of externalizing behaviors. In a follow up review, De Los Reyes and Kazdin (2005) noted that the absence of a theoretical framework to guide research on the accuracy of discrepancy assessment ratings research was a large limitation in research. In addition, the authors discussed how the motivations behind the ratings as well as the perception of abnormal or appropriate behavior could also cause discrepancies between informants. Their 2005 review addressed types

of characteristics that could impact the ratings of an informant (parental stress, parental acceptance, family SES, child age, etc.), in addition to the discrepancies of rater pairs.

A pattern is commonly seen throughout literature in the last decade, which includes a lack of teacher pairs and a lack of reasoning behind the consistent findings of rating discrepancies (De Los Reyes & Kazdin, 2005). The prevalence of research on ADHD has increased greatly over the last three decades, but very little has focused on assessment beyond symptoms, such as the best way for information from method and informant ratings to be paralleled (Pelham et al., 2005). These inconsistencies would be reflected through ratings obtained from sources who observe behaviors in the same settings, which provides a measure of interrater reliability (Bergeron et al., 2008). Interrater reliability, or the relation among each individual's ratings and scores, is a consistent source of information used for examining assessments. Studies examining interrater reliability from raters in the same settings have typically generated moderate correlation coefficients (.50 to .70; Mattison, Gadow, Sprafkin, Nolan, & Schneider, 2003), which could likely be due to a difference of rater's perceptions of behaviors and memory recall (De Los Reyes & Kazdin, 2005). De Los Reyes and Kazdin (2005) reported that prior research has also "examined informant discrepancies with regard to specific externalizing problems, such as aggression, hyperactivity/inattention, and oppositional behavior… and with regard to informants' ratings of childhood hyperactivity/inattention, prior work has generally revealed low-to-moderate levels of informant agreement" (p. 487). Furthermore, Pelham et al. (2005) reviewed rating scales for ADHD symptoms and summarized the reliability and validity information for each scale included in the systematic review and found that when cross-informants' ratings result in low reliabilities, the raters have differing evaluations of ADHD behavior. Pelham et al. (2005) reported interrater reliability coefficients

ranging from .63 to .69 for teachers on the BASC-2 Teacher Rating Scale Attention Problems scale (Reynolds & Kamphaus, 2004). In addition, Pelham et al. also reported a value of .61 for interrater reliability between teachers on the ASEBA CBCL (Achenbach & Rescorla, 2001).

Limitations of Previous Behavior Rating Scale Research

Instrument, source, and setting variance are generally evaluated using classical test theory analysis. This type of analysis yields valuable information, but in each, the source of error variance is evaluated individually—therefore not providing a broad view of an instrument's consistency across contexts in which it might be used. Generalizability (G) theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), however, allows researchers to concurrently examine multiple sources of error (across instruments and across raters), in addition to revealing the ways measurement error can be reduced. G theory offers increased usefulness for assessment research, as it extends the concept of measurement error beyond classical test theory, which cannot simultaneously evaluate multiple sources of error variance.

Researchers have applied G theory to evaluate sources of error in systematic direct behavior observation (Hintze & Matthews, 2004), behavior rating scales completed after classroom observations (Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese, 2007), and parent and teacher behavior rating scales (Briesch, Chafouleas, & Johnson, 2016; Wright & Piersel, 1992). G Theory allows these sources of error to be estimated concurrently, which is accomplished through the estimation of variance components. For example, Briesch et al. (2016) reviewed a variety of published generalizability theory studies and explained how variance components can concurrently compare different sources of error while also investigating how the sources of error connect through interactions. However, only Bergeron et al. (2008) examined specific sources of error within a G theory framework with teacher rating scales.

Bergeron et al. (2008) investigated the reliability and dependability of externalizing behaviors of children represented via rating scales completed by teachers. The study used two rating scales, the Behavioral Assessment System for Children, Second Edition Teacher Rating Scale-Child (Reynolds & Kamphaus, 2004) and the Achenbach System of Empirically Based Assessment, Teacher Report Form for Ages 6 to 18 (Achenbach & Rescorla, 2001). Teacher pairs completed both rating scales twice across a three-week period. Pearson correlation coefficients and G theory analyses were used to examine consistency between raters, rating scales, and measurement occasions. Using classical test theory, Bergeron et al. reported test-retest correlation coefficients for externalizing composites at or near the .90 criterion, which implies consistency over brief periods of time with teacher ratings, scores, and student behaviors. The G theory analyses results indicated that 8% to 17% of the total variance was due to the instrument facet, 4% to 5% was due to rater facet, and the time facet contributed only negligible levels of variance.

Thus, Bergeron et al. (2008) found that the largest source of error variance was attributed to teachers rating students' aggressive behavior at approximately 10-12%. Systematic differences between the two rating scales were reported as the next largest percentage of error variance at 8% of the total variance. In addition, the study reported variance contributed by the rating scales as the largest proportion of error at approximately 16% for each analysis section. This influence factored in the production of differing scores for students' conduct problems. However, limitations exist in the previously mentioned behavior rating scales. Both the BASC-2 (Reynolds & Kamphaus, 2004) and the ASEBA CBCL (Achenbach & Rescorla, 2001) were collected on clinical samples with high rates of ADHD (Pelham et al., 2005) for reliability and validity, though there were no ADHD subscales reported for all assessments. In addition, the

Technical Manual for the Conners Rating Scale (Conners et al., 1998) did not report interrater reliability with teachers, which would show consistency within the Teacher Rating Form.

Purpose of the Study

The purpose of this study was to examine targeted individual differences in the display of ADHD symptoms and error variance from rating forms completed by teachers across rater pairs and forms using Generalizability (G) theory. Because the ADHD Rating Scales are frequently used in assessing and diagnosing children with ADHD but often only evaluated by using a classical test theory approach, examination of scale properties is limited to examining only one source of variance at a time. Although each classical test theory method (e.g., test–retest or interrater reliability) provides valuable information, none provides a true comprehensive picture of the reliability of an instrument as each variance source is examined in isolation (Bergeron et al., 2008; Shavelson & Webb, 1991).

The study used G theory in determining what proportion of variance could be attributed to which of the study's focus on two sources of variance—rater and form. The approach outlined by Bergeron et al. (2008) provides the foundation and greatest influence for this comparison replication study using the ADHD Rating Scale School Versions. This scale was chosen due to its recent update and release, the absence of studies evaluating this new scale, and its widely used and perceived trusted reputation. Thus, this study investigated the reliability of the ADHD Rating Scale-IV: School Version (DuPaul et al., 1998), based on the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition-Text Revision* (DSM-4-TR; APA, 2000), and the ADHD Rating Scale-5, School Version (DuPaul et al., 2016), based on the DSM-5 (APA, 2013). Similar to Bergeron et al. (2008), error variance components and their interactions and absolute dependability coefficients, that reflect the degree to which a single score from a

scale can be generalized to scores of other scales measuring the same behavioral domain, were calculated. These analyses were based on the subscale and total scores for both the ADHD Rating Scale-IV (DuPaul et al., 1998) and the ADHD Rating Scale-5, which measure Inattention and Hyperactivity/Impulsivity (DuPaul et al, 2015). It was hoped that the information gathered might continue to benefit others in better understanding the reliability and generalizability of the ADHD Rating Scale scores used in the process of assessing and diagnosing students with ADHD.

To summarize, the aims of this study were to investigate the reliability of the ADHD Rating Scale-IV and the ADHD Rating Scale-5, to replicate the Bergeron et al. (2008) study with the ADHD Ratings Scales addressing source and instrument variance, and to use the recently released ADHD Rating Scale-5 in a comparison replication study.

Methods

Participants

Seven teacher pairs, or 14 total teachers, were recruited to participate in the study. Due to attrition, data were collected on 6 dyads of co-teachers (12 total teachers) from two local elementary schools in Memphis, Tennessee. This represents about half of the teacher pairs of classrooms with children age 7-10 at these two schools. These elementary schools were selected to keep a consistent class structure with a single primary classroom teacher for the included grades (grades 2-4) in the study, as well as to represent urban public and suburban private environments. These teachers are paired for academics as the students move between the two classrooms for academic subjects daily. In addition, these teachers are paired for school activities, such as lunch and recess. An effort was made to secure participants from the same grade levels across schools, and resulted in one second grade classroom and four third grade

classrooms. Although the ADHD scales are normed for participants ranging between 5 and 10 years of age, the setup of the classrooms for the lower end of the ages prohibited recruiting 1st grade classrooms because they do not have paired teachers.

The secondary participants were the students in grades two through four (and ages 8 to 10) who were in the co-teachers' classroom and were randomly selected to be independently rated by the co-teachers. These students were not involved in the actual process of the study, nor were they identified at any time during the data collection. Only their gender, race, age in years, and grade level were reported by the co-teachers. IRB approval was obtained prior to the initial contact with the participants and the school principals. (Appendix F).

Measures

ADHD Rating Scale-IV: School Version (DuPaul et al., 1998). The ADHD Rating Scale-IV: School Version is a behavior rating scale that lists the nine inattention and nine hyperactivity/impulsivity symptoms of ADHD from the DSM-IV (APA, 2000). Teachers rate the target child's behaviors on a scale from 0 (*never or rarely*) to 3 (*very often*) for each symptom item. The measure yields a total score, as well as subscale scores for inattention and hyperactivity/impulsivity in both raw score and percentile rank form. Items from the Inattention scale include "loses things necessary for tasks or activities" and "does not seem to listen when spoken to directly." Items from the Hyperactivity/Impulsivity scale include "fidgets with hands or feet or squirms in seat" and "has difficulty awaiting turn." This study employs percentile ranks, based on age- and gender-based norms (girls and boys ages 8 to 10), from both subscales as the primary variables of interest.

The norming sample for the ADHD Rating Scale-IV: School Version consisted of 4,009 randomly selected students rated by their teachers (n = 2,005). Each teacher rated one randomly

selected boy and one randomly selected girl from his or her classroom roster to ensure equal gender representation. Teachers were predominately Caucasian (90.6%), had a mean of 14.6 years teaching experience, and were recruited from 31 school districts across the United States. The norming sample included ratings from both teachers who instructed general educational courses (83.3%) and special education courses (16.4%). Teachers taught the full grade range (kindergarten through grade 12). The districts from which teachers were drawn were selected based on the 1990 U.S. Census, grouped by region, and selected by stratified random sampling.

These teachers rated a total of 4,009 students (1,040 boys, 945 girls, and 12 unspecified) who were selected to approximate the U.S. Census (1990) data distributions for ethnic group and geographic region of the U.S. They ranged in age from 4 to 19 years, and attended kindergarten through 12th grade. The majority of the sample was White non-Latino (65.1%), with African-American being the other well-represented group (18.5%). Boys were reported to exhibit more frequent inattentive and hyperactive-impulsive behaviors than girls, younger children received high ratings of ADHD symptoms than older children, and children who were African-American children were rated by teachers to exhibit more frequent ADHD-related behaviors than were children who were Caucasian and Latino. The normative data for the ADHD Rating Scale-IV is representative of the U.S. population (1990 census) in respect to ethnic group.

For the ADHD Rating Scale-IV: School Version, coefficient alpha values were calculated to determine the internal consistency of the composites based on the entire norming sample, which were .96 for the Inattention subscale and .88 for Hyperactivity/Impulsivity subscale. For test-retest reliability, teacher ratings were available for 52 children whose behavior was rated twice over a 1-month period between May and June 1995. Resulting test-retest reliability coefficients were as follows: Inattention = .89 and Hyperactivity-Impulsivity = .88. Interrater

reliability analysis employing teachers (who completed the ADHD Rating Scale-IV– School Version) and parents (who complete the ADHD Rating Scale-IV– Home Version) produced Pearson product-moment correlation coefficients of .45 for Inattention and .40 for Hyperactivity-Impulsivity. Although interrater reliability (or agreement) analysis employing only teachers using the ADHD Rating Scale-IV– School Version has not been reported in the literature, Pelham et al. (2005) demonstrated comparable internal consistency and test-retest reliability results to those reported by DuPaul et al. (1998).

Several types of validity evidence were provided by DuPaul et al. (1998). When testing for evidence of internal relations, both an exploratory factor analysis and a confirmatory factor analysis of items were conducted. Results from both analyses showed that either a 1-factor or a 2-factor model demonstrated acceptable fit. Consistent with the organization of the ADHD symptoms into two categories in the DSM-IV, the 2-factor model was chosen as the model for organizing the scale.

Additional validity evidence came from students, who ranged from kindergarten to grade 8, rated by teachers using the ADHD Rating Scale-IV: School Version and criterion measures. The strongest correlations were found between the Conners Teacher Rating Scale-39 (CTRS-39; Conners, 1989) Hyperactivity and Hyperactivity Index scores and the ADHD Rating Scale-IV: School Version Hyperactivity/Impulsivity subscale scores. The ADHD Rating Scale-IV: School Version Inattention subscale score was also correlated significantly with the CTRS-39 Daydream-Attention scale.

After examining samples with the clinical condition of ADHD, DuPaul et al. (1998) reported statistically significant differences between control groups and Inattentive subtype groups on the ADHD Rating Scale-IV: School Version for both Inattention and

Hyperactivity/Impulsivity. The Hyperactivity-Impulsivity scale was also significantly higher for participants in the Combined subtype group. After these group differences were found, the results of logistic regression analyses were used to evaluate the predictive validity. The analyses for the Inattention subscale indicated teachers were able to differentiate children with ADHD Inattentive type from control groups and children with ADHD Combined type from controls; analyses of the Hyperactivity-Impulsivity subscale indicated that it successfully differentiated children with ADHD combined type from those with ADHD Inattentive type. These results indicated that teacher ratings on the ADHD Rating Scale-IV are extremely important in predicting subtype membership. Pelham et al. (2005) also found that the ADHD Rating Scale-IV School Version yielded significant differences between ADHD and control groups in distinguishing among subtypes of ADHD.

ADHD Rating Scale-5, School Version: Child Form (DuPaul et al., 2016)**.** The ADHD Rating Scale-5 is a behavior rating scale that lists nine inattention and nine hyperactivity/impulsivity symptoms of ADHD from the DSM-5 (APA, 2013) that are rated on a 0 (*never or rarely*) to 3 (*very often*) scale. The measure yields a total score, as well as subscale scores for inattention and hyperactivity/impulsivity in both raw score and percentile rank form. The measure uses a Child Form (targeting children ages 5 to 10) and an Adolescent Form (targeting adolescents ages 11 to 17), and this study employed the Child Form. Items for the Inattention scale include "avoids, dislikes, or is reluctant to engage in tasks that require sustained mental effort" and "forgetful in daily activities." Items for the Hyperactivity/Impulsivity scale include "interrupts or intrudes on others" and "has difficulty waiting his or her turn." The scale also includes items representing six domains of impairment (i.e., relationships with significant others, peer relationships, academic functioning, behavioral functioning, homework functioning,

and self-esteem) rated once following the inattention items and again after rating the hyperactivity-impulsivity items. Items are prefaced with "How much do the nine behaviors in the previous question cause problems for this child?" and include "controlling behavior in school" and "feeling good about himself/herself." This study employed percentile ranks, based on age- and gender-based norms (girls and boys ages 8 to 10), from the Inattention and Hyperactivity/Impulsivity subscales as the primary variables of interest. The information from the impairment scale items was not analyzed.

The norming sample for the ADHD Rating Scale-5, School Version consisted of 1,070 teachers who provided ratings. Each teacher rated one randomly selected male and one randomly selected female from his or her classroom roster to ensure equal gender representation. Teachers were predominately white non-Hispanic (87.3%), had a mean of 17.9 years teaching experience, were recruited from all regions of the United States, and included teachers for kindergarten through Grade 12. The teachers were recruited through GfK and e-Rewards, which are two national research firms. To be included, teacher must have indicated employment as a full-time teacher, which excluded those who are substitute teachers. A total of 1,399 teachers completed ratings, with 596 qualifying for inclusion based on their student's demographics, which were child grade, race, ethnicity, and geographic region.

The 2,140 students who were rated by teachers included 1,040 males and 1,070 females who ranged in age from 5 to 17 years of age (M = 11.53) and attended kindergarten through grade 12. Most students were White non-Hispanic (54.8%), Black non-Hispanic (12.7%), other non-Hispanic 7.0%), Hispanic (24%), or biracial non-Hispanic (1.5%) backgrounds. As part of ADHD Rating Scale-5 norming, 1,027 Child Form ratings were completed, including 963 in the School Version. The male sample participants by age group for the Child Form were 238 for age

5 to 7 and 253 for age 8 to 10. The female sample participants by age group for the Child Form were 228 for age 5 to 7 and 244 for age 8 to 10.

Coefficient alpha values were calculated to determine the internal consistency of the composites based on the norming sample, which were .96 for the Inattention subscale and .95 for Hyperactivity/Impulsivity subscale for the ADHD Rating Scale-5. Internal consistency analyses were not conducted for impairment scores. Test-retest reliability was analyzed using a group of 64 children (37 boys, 27 girls) ranging in age from 5-17 years who attended kindergarten through 12th grade; teachers completed the scales approximately 6 weeks apart. Test-retest reliability coefficients were .91 for Inattention and .90 for Hyperactivity/Impulsivity.

Though no interrater agreement analysis employing only teachers using the ADHD Rating Scale-5, School version has been completed or reported in the literature, DuPaul et al. (2016) reported several other types of validity evidence. In terms of internal relations, they examined the fit of the correlated 2-factor structure of the ADHD Rating Scale-5 using data from a sample of 1,070 teachers. This structure conformed with the DSM-5 diagnostic criteria and is consistent with the scale's Inattention and Hyperactivity/Impulsivity subscales; thus, it was chosen over the 3-factor model. This 2-factor structure was well fitting and invariant across child gender, age, informant, informant gender, and language.

In terms of relations with other variables, DuPaul et al. (2016) examined classroom behavioral observations, academic performance, and the Conners-3 Teacher Rating Scale (Conners, 2008) as factors for criterion-related validity evidence. The Conners-3 Teacher Rating Scale (CTRS, 2008) Inattentive Scale score had a correlation coefficient of .85 with ADHD Rating Scale-5 (2016) Inattention subscale and .75 with Hyperactivity-Impulsivity, and the CTRS Hyperactivity Scale score had a correlation coefficient of .77 with the Inattention subscale

and .89 with the Hyperactivity-Impulsivity subscale scores from the ADHD Rating Scale-5. All Pearson correlation coefficients were statistically significant, ranging from .46 to .89. No other comparison studies that used only scores from the ADHD Rating Scale-5 or the School Version or Home Version have appeared in the literature. Table 1 summarizes the reliability estimates from the respective manuals.

Table 1
Reliability Estimates as Reported in the Manuals for the ADHD Rating Scale-IV and ADHD Rating Scale-5, School Versions

	Internal Consistency	**Test-Retest**
ADHD Rating Scale-IV		
Inattention	.96	.89
Hyperactivity/Impulsivity	.88	.88
ADHD Rating Scale-5		
Inattention	.96	.91
Hyperactivity/Impulsivity	.95	.90

Note. Information obtained from DuPaul et al. (1998) and DuPaul et al. (2016).

Adaptive Behavior Assessment System: Teacher Form (ABAS; Harrison & Oakland, 2000). The Adaptive Behavior Assessment System Teacher Form is a behavior rating scale assessing the adaptive skills of individuals who are school aged. For this study, only the Leisure subscale was completed. Items from the leisure scale include "participates with others in a game or other activity without needing encouragement" and "waits for his/her turn in games and other classroom activities." Teachers are asked to rate the behavior frequency on a 4-point scale (0 = *is not able* to 3 = *always when needed*). This scale was selected to produce interference between the co-teachers' completion of the two versions of the ADHD Rating Scales – School Version. The brief number of (n = 17) and contrasting adaptive behavior related questions of the Leisure scale were factors in the selection of this scale. For this study, the ABAS was used to break-up the administration of the ADHD-IV and ADHD-5 scales and was not analyzed further.

The development and norming of the ABAS, and reliability and validity evidence supporting the use of the Leisure subscale were reported by Harrison and Oakland (2000). For example, internal consistency coefficient alpha values for Leisure from the ABAS the norming sample was .94 for age 8, .93 for age 9, and .93 for age 10.

Procedures

After consulting with relevant school officials (e.g., principals) and identifying classrooms with teacher dyads, a letter of invitation within the informed consent form (Appendix A) was delivered to teachers of second-, third-, and fourth-grade classrooms. These letters also included an information sheet for teachers to complete (Appendix B). Teachers who were willing to participate were asked to sign the informed consent form. Two copies of the informed consent form were included in the envelope for each teacher—one to sign and return and another to keep. Consent from both teachers within a classroom was obtained before the classroom was enrolled in this study. In addition, demographic information (Appendix B) from each teacher was collected. Along with information on their school and grade taught, the information included race, ethnicity, age, gender, and years of education. This information was used to describe the participants and may be used to outline potential influences on collected data for future studies.

Teacher pairs were provided a list of 10 numbers, selected randomly from the total number of students they reported (using a random number generator) on their class roster, and were asked to sort their class roster alphabetically by students' last names. The teachers then selected the 10 students associated with these 10 numbers provided by the investigators (with the number 1 associated with first name in alphabetical order), which represented approximately half of their class roster. This list of randomly selected 10 numbers (Appendix C) was also linked to

the participant ID numbers used to mark each rating packet and other study materials. The participant ID included indicators associated with the primary participants and the school in which they were enrolled. Researchers did not have contact with students. From the perspective of the investigator, these participant ID numbers were the sole identification for students for the duration of the study, and the students were the secondary participants.

Next, both members of the teacher dyads were given three brief behavior rating scales each and a few questions per randomly selected student. The rating scales were delivered in packets pre-assigned to each participating teacher. The lead investigator delivered the scales to all schools the same day and explained the procedures to each teacher. Every teacher had the opportunity to ask questions at that time. The investigator picked up the completed and sealed packets at the end of the same week. The teachers were asked to complete all standardized items on the scales (a total of 64 items per student, which were estimated to take about 5 min per student), with the exception of the impairment ratings on the ADHD Rating Scale-5, which are not normed. The two ADHD Rating Scales were administered in a counterbalanced order; the Leisure subscale of the ABAS Teacher Form was always completed second in sequence. Half of the teacher dyads completed the ADHD Rating Scale-IV first, and half of the dyads completed the ADHD Rating Scale-5 first. Each teacher was asked to fold and seal the first ADHD Rating Scale before beginning the second scale for each child. In addition, each teacher was asked to fill out a questionnaire (Appendix D and Appendix E) including items about the independent completion of the ratings to ensure the integrity of responses within teacher dyads, as the teachers were asked not to communicate with their teaching partner about ratings of specific students.

Analyses

The focus of the study was to examine sources of variance in obtained scores from teacher-completed rating scales of children's behavior and, more specifically, to determine the generalizability of both the ADHD Rating Scale IV: School Version and ADHD Rating Scale-5: School Version ratings across the facets of rater and form. Scores from both scales were converted to percentile scores for analysis. Pearson correlations were used to examine relations among scores from the total ADHD scores, and the Inattention and Hyperactivity/Impulsivity subscales across forms of the instrument (ADHD Rating Scale-IV and ADHD Rating Scale-5) and across teacher rater (rater 1 and rater 2).

IBM-SPSS 24 was used to determine the proportion of variance attributable to sources of error variance (rater and form), their interactions, and residual variance for the Inattention and Hyperactivity/Impulsivity subscales (see Appendix E). This G theory study used a univariate format where all of the facets were random, and the design was partially nested where the students and raters were nested within the classrooms. Variance components were computed for the following sources of variance: 1) main sources included classroom, student-within-classroom, rater-within-classroom, instrument; 2) interactions included student-by-rater-within-classroom, classroom-by-instrument; and 3) residual error. Absolute dependability coefficients, which reflect the extent to which a single score from a scale can be generalized to other scales' scores measuring the same behavioral domain across rater, were also computed for the following pairs of scores: 1) percentile scores for ADHD-IV total and ADHD-5 total; 2) inattention ADHD-IV and inattention ADHD-5 subscales; and 3) Hyperactivity/Impulsivity ADHD-IV and Hyperactivity/Impulsivity ADHD-5 subscales.

Results

Due to unresponsiveness, data were received for 6 of the targeted 7 teacher pairs. While analyzing the data, an additional teacher pair was determined to have incorrectly completed the packets, and was not included in analysis. Therefore, 5 pairs of teachers and 50 students were used for the analyses. These teachers were all self-identified as female, with 50% (5 teachers) identifying as African American and 50% (5 teachers) identifying as Caucasian. The average total amount of years teaching across the teachers was 14.6 years, ranging from 3 years to 33 years of experience, while the average age of teachers was 38.2 years, ranging from 26 to 55 years of age.

According to the teachers, the students, or secondary participants (n = 50), had an average age of 8.7 years and were in grades 2 and 3, as the grade 4 participants were identified as inadequate due to packet incompletion. Of the 50 participants, 54% were identified as male, 33% were identified as female, and 13% were missing gender identification. For race, 44% were identified as African-American, 30% were identified as Caucasian, 17% for Hispanic, 3% for Asian, 1% for Biracial, and 4% for other (race was not reported for 1 student).

The means, standard deviations, and ranges for the inattention subscale, hyperactive/impulsive subscale, and total scores for both the ADHD Rating Scale-IV, School Version and the ADHD Rating Scale-5, School Version, Child Form are seen in Table 2 in both percentile ranks and raw score form. Across the percentile scores, the means of the Inattention subscales were higher (41.98 for ADHD Rating Scale-IV and 38.74 for ADHD Rating Scale-5) than the means of the Hyperactivity/Impulsivity percentile scores (28.44 for ADHD Rating Scale-IV and 37.02 for ADHD Rating Scale-5). The total percentile score for the ADHD Rating Scale-IV, School Version (38.98) was slightly lower than the ADHD Rating Scale-5, School

Version (39.76). The scores were varied across the ADHD characteristics for clinical significance which indicated that we had a wide-range of scores where some participants met the clinical qualifications for potential diagnosis, but the majority did not.

Table 2
Means, Standard Deviations, and Ranges Across Measurements

Name	*Mean* *%ile (score)*	*SD* *%ile (score)*	Range *%ile (score)*
ADHD Rating Scale-IV			
Inattention	41.98 (8.27)	30.43 (8.43)	1 to 98 (0 to 27)
Hyperactivity/Impulsivity	28.44 (5.37)	33.31 (7.60)	1 to 95 (0 to 27)
Total Score	38.98 (13.64)	30.84 (4.88)	1 to 98 (0 to 54)
ADHD Rating Scale-5			
Inattention	38.74 (7.53)	29.76 (7.86)	1 to 95 (0 to 27)
Hyperactivity/Impulsivity	37.02 (6.11)	31.35 (7.28)	1 to 99 (0 in 27)
Total Score	39.76 (13.64)	30.45 (4.88)	1 to 99 (0 to 54)

Note. ADHD Rating Scale-IV = ADHD Rating Scale-IV, School Version. ADHD Rating Scale-5 = ADHD Rating Scale-5, School Version, Child Form.

Correlations

Pearson correlations were computed to examine associations between raters and instruments, including percentile scores obtained from different raters on the same instrument and different instruments from the same rater. These correlation coefficients can be seen in Table 3. The instrument correlation between the Inattention subscales was .92, with the instrument correlation between Hyperactivity/Impulsivity subscales at .85. Correlations between 'A' Raters, or those teachers who completed the ADHD Rating Scale-IV first in order on each of their 10 students, were .91 for Inattention and .86 for Hyperactivity/Impulsivity. The correlations for 'B' raters, or those who completed the ADHD Rating Scale-5 first in order on each of their 10 students, were .94 for Inattention subscales and .82 for Hyperactivity/Impulsivity

subscales. The correlations for Total Score across instruments, A order raters, and B order raters were .98, .99, and .97, respectively.

All correlations in Table 3 were statistically significant at $p < .01$. In addition, correlations between instruments and between raters were all strong to very strong, ranging from .82 to .99.

Table 3
Correlations Between Raters and Instruments

		Rater		
ADHD Rating Scale-IV	ADHD Rating Scale-5	A	B	Instrument
Inattention	Inattention	.91*	.94*	.92*
Hyperactivity/Impulsivity	Hyperactivity/Impulsivity	.86*	.82*	.85*
Total Score	Total Score	.99*	.97*	.98*

* Correlations are significant at the 0.01 level

Variance Components

The variance estimates, which can be seen in Table 4, show the variance component estimates and the dependability coefficients across score comparisons across the three components of Hyperactivity/Impulsivity, Inattention, and Total Score.

Table 4
Variance Component Estimates Among Facets by Score Comparison

Facets	Estimated Variance Component		
	Hyperactivity/Impulsivity	*Inattention*	*ADHD Total*
Classroom	0.0 (0%)	18.04 (1.8%)	0.0 (0%)
Student within Classroom	351.45 (26.8%)	119.25 (11.7%)	210.38 (19.8%)
Rater within Classroom	228.87 (17.5%)	106.60 (10.4%)	0.0 (0%)
Instrument	10.43 (0.8%)	0.0 (0%)	165.40 (15.5%)
Student-by-rater within Classroom	528.76 (40.3%)	698.16 (68.3%)	668.08 (62.7%)
Classroom by instrument	14.92 (1.1%)	7.33 (0.7%)	0.06 (0.01%)
Residual	176.95 (13.5%)	73.45 (7.2%)	21.56 (2.0%)
Total	1311.37	1022.84	1065.49
ϕ	.96	.91	.97

Note. Proportion of total variance in parentheses. Negative estimated variance components were set to zero.

Hyperactivity/Impulsivity. For the Hyperactivity/Impulsivity subscale, more than one third of the total variance (40.3%) was due to differences of teachers' evaluation of a student's behavior within each classroom. Approximately one fourth (26.8%) of the variance among Hyperactivity/Impulsivity estimates was due to differences in a child's behavior, rather than error. In addition, 17.5% of variance was attributed to overall differences in the way the raters within each classroom approached the task of completing the rating scale. Residual variance accounted for an additional 13.5%, which could be random events or other facets not addressed. In addition, the dependability coefficient was .96, which is strong and slightly stronger than the Pearson correlation coefficients. Thus, differences were seen when looking at the facets separately versus simultaneously.

Inattention. For the Inattention component, over two thirds (68.3%) of the total variance was attributed to inconsistencies of the teachers' evaluation of a particular student's behavior within each classroom. Therefore, teachers in some classrooms rated certain students' behaviors differently, while other teachers consistently marked higher ratings than her teaching partner. The estimated variance due to differences in a child's behavior was 11.7%, and an additional 10.4% was attributable to overall differences in the way raters approached the task of completing the rating scale. In addition, 7.2% of the total variance was due to residual variance. The dependability coefficient for the Inattention component, which looks at the facets simultaneously, was .91, which shows strong dependability but is slightly lower than the Pearson correlation coefficients.

ADHD Total Scores. Results were also similar for the Total Score component. Over half of the total variance (62.7%) was attributed to the student-by-rater within classroom, or the inconsistencies of teachers' ratings for the overall ADHD characteristics of a student. The differences in students' behavior, instead of error, accounted for 19.8% of the total variance, and 15.5% was attributable to instrument variance. This could be contributed to by differences in scores that the ADHD Rating Scale-IV and the ADHD Rating Scale-5 produce, due to differences in updated wording. Two percent (2%) of the variance was attributable to residual variance, or random events. In addition, the dependability coefficient for the Total Score component was .97, which indicates a strong dependability yet slightly weaker than the Pearson correlations.

Discussion

This study used G Theory to examine what proportion of variance could be attributed to two main sources of variance, rater and form, simultaneously instead of only looking at the

facets in isolation for two different instruments that assess the same constructs. The ADHD Rating Scale-IV, School Version and the ADHD Rating Scale-5, School Version assess total ADHD ratings, Hyperactivity/Impulsivity, and Inattention subscales. Results showed that the largest sources of variance across all three components was attributed to teachers rating the behaviors for Hyperactivity/Impulsivity, Inattention, and combined behaviors for Total Score differently than their teaching partner had rated the same student. The second largest percentage for all three components was the student within classroom facet, which is due to individual differences in student behavior rather than a source of error. The dependability coefficient for all three components was very strong, ranging from .91 to .97. Thus, as indicated by Cronbach and colleagues (1972), for this study the G theory offered increased usefulness for assessment research of the ADHD rating scales as it extended the concept of measurement error beyond the classical test theory, and simultaneously evaluated multiple sources of error variance.

Instrument Variance

All Pearson correlation coefficients between corresponding scales from both the ADHD Rating Scale-IV, School Version and the ADHD Rating Scale-5, School Version were considered strong. Previous research that examined consistency between comparable scores from different assessments has reported similar results (Bergeron et al., 2008; Mattison et al., 2003). Overall, the dependability coefficients were stronger than the Pearson correlations that were obtained from the scale technical manuals and presented in Table 1. Thus, scores demonstrated strong reliability even when examined with the effect of multiple variance sources, which indicates how a score can be generalized to others that measure the same area from different raters and instruments (Shavelson & Webb, 1991). Results of the G theory analyses explained that the instrument main effect contributed between 0% and 15.5% across the

constructs, which is a lower range than the findings of Bergeron et al. (2008). One possible explanation for this is that the instruments were more similar than in the Bergeron study and used like methods for obtaining percentile scores. The results of this study also show that the ADHD Rating Scale-5 meets the reliability coefficient criteria of .90 or higher to be used for significant decisions, such as school placement, educational qualification, or diagnoses.

Source Variance

The student-by-rater within classroom facet accounted for the largest proportion of total variance across all scores. These are desirable findings, as they suggest the leading contributors to differences in scores are the differences in student behaviors or the teacher interpretation of student behaviors and not the instrument itself. In addition, this is expected as prior research has shown low-to-moderate levels of agreement between informants (De Los Reyes & Kazdin, 2005), and teachers often have discrepancies in ratings on a scale due to response bias (Chi & Hinshaw, 2002; De Los Reyes & Kazdin, 2004; Youngstrom et al., 2000). Bergeron et al. (2008) also found that the source of error variance was due to teacher ratings student behavior, with the largest proportion of error variance attributed to the students within the classroom. This study found a lower percentage of instrument variance, with larger source variance.

However, the large percentage of variance estimates influenced the obtained scores. This could be due to a difference in rater perception of behaviors (De Los Reyes & Kazdin, 2005), or to the different amounts of time raters in two of the five teacher dyads spent with the rated students. These teachers reported themselves as partners, as their students moved between the two classrooms for educational lessons and were partnered for times such as lunch and research together. However, these teachers did not spend equal amounts of time each day with the

partnered classroom, and the main teacher would have additional opportunities to observe her classroom students during transitions and additional activities.

Limitations

Although this study provides useful information about the dependability of the recently revised ADHD Rating Scale-5, results of the study do have limitations. First, this study focused on elementary students who were 7 to 10 years of age. Because the age guidelines for the Child Form on the ADHD Rating Scale-5 which range from ages 5-10, this study is most applicable to a more narrow range of students (ages 7-10) and would only generalize to that age group. A wider age range would be ideal in gathering a more representative sample from the specified population, as well as take into account the influence of developmental factors that vary considerably between the ages of 5 and 10. In addition, this study took place in the southeastern region of the United States. Although the two selected schools differed in setting, academic curriculum, and type of enrollment, the geographical location of the two schools limits the ability to generalize the findings. Additional research with a larger geographic participation at the classroom- and student-level would be needed to better generalize the findings, as well as a larger number of teacher pairs.

Another limitation of the study is that the sample size had to be reduced to 5 teacher-pairings due to reporting inconsistencies of gender, race, and age by a set of teachers and a resulting loss of the 10 corresponding student participants that could not be included in the analyses. In order to ensure the students remained anonymous and the teachers rated the same student for each assigned ID number, the teacher pairs would need to compare their ratings for the corresponding student. This would have increased the number of teacher pairs and perhaps strengthened the resulting analyses.

In addition, this study focused on two facets, rater and instrument, as sources of variance. The addition of a time facet to the study would address the current limit this produces, and would improve the ability to generalize the findings to all sources of variance. The addition of time as a facet of variance would also improve the G-Theory analyses, as it would allow researchers to investigate the percentage of variance that temporal variance is responsible for, within the total percentage of variance.

Implications

This study was conducted as a way to assess the dependability of the ADHD Rating Scale-5 and learn more about the consistency of teacher ratings. Findings from this study will hopefully be beneficial to students in the future as it contributes to the accuracy of the ratings of their behaviors, as well as to teachers as it will communicate the consistency of teachers as they continue to complete behavior rating scales in the future.

In addition, school psychologists should consider the ratings of individual students by teachers cautiously (De Los Reyes & Kazdin, 2005). Findings from this study show that error variance in ratings by teachers can be attributed to their bias while rating certain students, which is consistent with previous findings (Bergeron et al., 2008; De Los Reyes & Kazdin, 2005). School psychologists should consider the competencies of the teacher rater and how those competencies can affect the ratings (De Los Reyes & Kazdin, 2005). Furthermore, this study confirms the importance of considering multiple informants and settings. Although these scales are dependable, multiple factors should be considered prior to a diagnosis of ADHD (Barkley, 2006). Just as errors of variance from teacher informants have been seen in this study, errors in the assessments given to the students as part of the information gathered during the evaluation are also seen. Consequently, although behavior rating scales are an excellent source of

information about the behaviors of a student, multiple sources of information should be obtained before a diagnosis is confirmed or denied. School psychologists should consult with teachers who present conflicting or differing ratings on students, as other factors such as classroom management could be influencing the behavior of a student in a classroom. These findings further highlight the need for the evaluation of academic functioning and classroom behavior to be comprehensive, especially in the evaluation of ADHD.

In addition, a future study could investigate the relationship between the results of the analyses with the demographic information collected from teachers, such as the years of experience, ethnicity, and age. Future studies could also examine the results between the urban public school and the suburban private school, as it could be beneficial for future research. Finally, a wider age range of secondary participants, or students, could give additional information within child development factors if included in future research in this area.

References

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

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

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Barkley, R. A. (2014). *Attention-deficit hyperactivity disorder: A handbook for diagnosis and treatment* (4th ed.). New York, NY: Guilford Press.

Barkley, R. A., Fischer, M., Smallish, L., & Fletcher, K. (2006). Young adult outcome of hyperactive children: Adaptive functioning in major life activities. *Journal of the American Academy of Child and Adolescent Psychiatry, 45,* 192–202.

Bergeron, R., Floyd, R. G., McCormack, A. C., & Farmer W. L. (2008). The generalizability of externalizing behavior composites and subscale scores across time, rater, and instrument. *School Psychology Review, 37*, 91-108.

Briesch, A. M., Chafouleas, S. M., & Johnson, J. (2016). Use of generalizability theory within K–12 school-based assessment: A critical review and analysis of the empirical literature, *Applied Measurement in Education, 29*, 83-107.

Chafouleas, S. M., Christ, T. J., Riley-Tillman, T. C., Briesch, A. M., & Chanese, J. M. (2007). Generalizability and dependability of daily behavior report cards to measure social behavior of preschoolers. *School Psychology Review, 36,* 63–80.

Chi, T. C. & Hinshaw, S. P. (2002). Mother-child relationships of children with ADHD: The role of maternal depressive symptoms and depression-related distortions. *Journal of Abnormal Child Psychology, 30*, 387-400.

Conners, C. (1989). *Conners rating scales: CPRS-39, CTRS-39*. Multi-Health Systems, North Tonawanda, NY.

Conners, C. K. (2008). *Conners rating scales* (3rd ed.). Toronto, ON, Canada: Multi- Health Systems.

Conners, C., Sitarenios, G., Parker, J. D., & Epstein, J.N. (1998). The revised *Conners Parent Rating Scale* (CPRS-R): factor structure, reliability, and criterion validity. *Journal of Abnormal Child Psychology, 26,* 257-268.

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

De Los Reyes, A., & Kazdin, A. E. (2004). Measuring informant discrepancies in clinical child research. *Psychological Assessment,* 16, 330-334.

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131,* 483–509.

DuPaul, G. J., Power, T. J., Anastopoulos, A. D., & Reid, R. (1998). *ADHD Rating Scale–IV: Checklists, norms, and clinical interpretation*. New York, NY: Guilford Press.

DuPaul, J. G., Power, T. J., Anastopoulos, A. D., & Reid, R. (2016). *ADHD Rating Scale-5 for Children and Adolescents: checklists, norms, and clinical interpretation.* New York, NY: Guilford Press.

DuPaul, G. J., Reid, R., Anastopoulos, A. D., Lambert, M. C., Watkins, M. W., & Power, T. J. (2015). Parent and teacher ratings of attention-deficit/hyperactivity disorder symptoms: Factor structure and normative data. *Psychological Assessment, 28,* 214-225.

Harrison, P. L., & Oakland, T. (2000). *Adaptive Behavior Assessment System*. San Antonio, TX: The Psychological Corporation.

Hintze, J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review*, *33*, 258–270.

Kamphaus, R. W., & Frick, P. J. (2002). *Clinical assessment of child and adolescent personality and behavior* (2nd ed.). Boston, MA: Allyn & Bacon.

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

Mattison, R. E., Gadow, K. D., Sprafkin, J., Nolan, E. E., & Schneider, J. (2003). A DSM–IV referenced teacher rating scale for use in clinical management. *Journal of the American Academy of Child and Adolescent Psychiatry*, *42*, 442–449.

Norfolk, P. A., & Floyd, R. G. (2015). Detecting parental deception using a behavior rating scale during assessment of attention-deficit/hyperactivity disorder: An experimental study. *Psychology in the Schools*, 52, 158-172.

Pelham, W. E., Fabiano, G. A., & Massetti, G. M. (2005). Evidence-based assessment of attention deficit hyperactivity disorder in children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, 34, 449-476.

Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior Assessment System for Children, Second*

Edition. Circle Pines, MN: American Guidance Service.

Sax, L. & Kautz, K. J. (2003). Who first suggests the diagnosis of attention deficit/hyperactivity disorder. *Annals of Family Medicine, 1*, 171-174.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Sibley, M.H., & Pelham, W. E. (2011). Diagnosing ADHD in adolescence. *Journal of Consulting and Clinical Psychology, 80*, 139-150.

Volpe, R. J., Briesch, A. M., & Gadow, K. D. (2011). The efficiency of behavior rating scales to assess inattentive-overactive and oppositional-defiant behaviors: Applying generalizability theory to streamline assessment. *Journal of School Psychology*, 49, 131-155

Whitcomb, S. A., & Merrell, K. W. (2012). *Behavioral, social, and emotional assessment of children and adolescents* (4th ed.). Mahwah, NJ: Erlbaum.

Wolraich, M. L., Lambert, W., Doffing, M. A., Bickman, L., Simmons, T., & Worley, K. (2002). Psychometric properties of the Vanderbilt ADHD diagnostic parent rating scale in a referred population. *Journal of Pediatric Psychology, 28*, 559-568.

Wright, D., & Piersel, W. C. (1992). Components of variance in behavior ratings from parents and teachers. *Journal of Psychoeducational Assessment*, *10*, 310–318.

Youngstrom, E., Loeber, R., & Stouthamer-Loeber, M. (2000). Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *Journal of Consulting & Clinical Psychology, 68*, 1038-1050.

APPENDIX A: Consent to Participate in a Research Study

Dependability of the Ratings Across Rater and Form

WHY ARE YOU BEING INVITED TO TAKE PART IN THIS RESEARCH?

You are being invited to take part in a research study about the psychometric properties of teacher-completed rating scales. You are being invited to take part in this research study because you are a part of a classroom that has two teachers working concurrently with students. If you volunteer to take part in this study, you will be one of about 12 teachers to do so.

WHO IS DOING THE STUDY?

The person in charge of this study is Meredith Manguno (*Lead Investigator, LI*), a doctoral student in the School Psychology program at the University of Memphis. She is being guided in this research by Dr. Michelle Stockton, School of Health Studies, and Dr. Frank Andrasik, Chair of Psychology Department. There may be other people on the research team assisting at different times during the study.

WHAT IS THE PURPOSE OF THIS STUDY?

We are looking at the way rating scales targeting student behavior are completed. It is important to note that we are not looking at the specific responses that either you or your teaching partner provide about any one child or your classroom as a whole. Rather, we are interested in and hope to learn more about the characteristics of the rating scales and how teachers complete them.

ARE THERE REASONS WHY YOU SHOULD NOT TAKE PART IN THIS STUDY?

We ask that you participate in this study because you are part of a classroom that has two teachers working concurrently with students. If you decide you cannot devote the time necessary to this study, we will understand.

WHERE IS THE STUDY GOING TO TAKE PLACE AND HOW LONG WILL IT LAST?

The research will be conducted at your school. The rating session will take about 50 minutes.

WHAT WILL YOU BE ASKED TO DO?

We will help you select 10 students (at random) in your classroom. Next, we will give you and your teaching partner each three brief behavior rating scales and a few questions per student and ask you to complete certain items on the scales (a total of 64 items per student, which we estimate will take about 6 minutes per student). We ask that you and your teaching partner not communicate about ratings of specific students. It is important to note that we are not looking at the specific responses that either you or your teaching partner provide. Rather, we are interested in characteristics of the rating scales.

WHAT ARE THE POSSIBLE RISKS AND DISCOMFORTS?

To the best of our knowledge, the ratings you will complete have no more risk of harm than you would experience each school day as a teacher.

WILL YOU BENEFIT FROM TAKING PART IN THIS STUDY?

There is no guarantee that you will benefit from taking part in this study. However, some teachers have experienced a better understanding of student behavior when completing rating scales and evaluating student behaviors. Your willingness to take part, however, may, in the future, help researchers and society as a whole better understand this research topic.

DO YOU HAVE TO TAKE PART IN THE STUDY?

If you decide to take part in the study, it should be because you really want to volunteer. You will not lose any benefits or rights you would normally have if you choose not to volunteer. You can stop at any time during the study and still keep the benefits and rights you had before volunteering.

IF YOU DON'T WANT TO TAKE PART IN THE STUDY, ARE THERE OTHER CHOICES?

If you do not want to be in the study, there are no other choices except not to take part in the study.

WHAT WILL IT COST YOU TO PARTICIPATE?

There are no costs associated with taking part in the study.

WILL YOU RECEIVE ANY REWARDS FOR TAKING PART IN THIS STUDY?

You will receive a $15 gift card for taking part in and completing this study.

WHO WILL SEE THE INFORMATION THAT YOU GIVE?

We will make every effort to keep private all research records that identify you to the extent allowed by law.

Your information will be combined with information from other teachers taking part in the study. When we write about the study to share it with others, we will write about the combined information we have gathered. You will not be personally identified in these written materials. We may publish the results of this study; however, we will keep your name and other identifying information private.

We will make every effort to prevent anyone who is not on the research team from knowing that you gave us information, or what that information is. The consent forms will be separated from teacher demographic information files, and the teacher demographic information will not be disclosed for any reason and all data will be kept in filing cabinets behind two locks.

We will keep private all research records that identify you to the extent allowed by law. However, there are some circumstances in which we may have to show your information to other people. For example, the law may require us to show your information to a court. Also, we may be required to show information which identifies you to people who need to be sure we have done the research correctly; these would be people from such organizations as the University of Memphis.

CAN YOUR TAKING PART IN THE STUDY END EARLY?

If you decide to take part in the study, you still have the right to decide at any time that you no longer want to continue. You will not be treated differently if you decide to stop taking part in the study.

The individuals conducting the study may need to withdraw you from the study. This may occur if you are not able to follow the directions they give you, if they find that your being in the study is more risk than benefit to you, or if the agency funding the study decides to stop the study early for a variety of scientific reasons.

WHAT IF YOU HAVE QUESTIONS, SUGGESTIONS, CONCERNS, OR COMPLAINTS?

Before you decide whether to accept this invitation to take part in the study, please ask any questions that might come to mind now. It would be helpful to list your questions at the bottom of this form. Later, if you have questions, suggestions, concerns, or complaints about the study, you can contact the investigator, Meredith Manguno at msmnguno@memphis.edu, or the faculty supervisors, Dr. Michelle Stockton at mstocktn@memphis.edu or Dr. Frank Andrasik at fndrasik@memphis.edu. If you have any questions about your rights as a volunteer in this research, contact the Institutional Review Board staff at the University of Memphis at 901-678-2705. We will give you a signed copy of this consent form to take with you.

_____ _____
Signature of person agreeing to take part in the study Date

_____
Printed name of person agreeing to take part in the study

_____ _____
Name of [authorized] person obtaining informed consent Date

Appendix B: Teacher Information Sheet

Please Tell Us <u>About Yourself</u> by
Completing the Blanks or by Placing Checks in the Boxes

Setting: General education
Special education

Grade taught: _____

Total number of years teaching: _____

Number of years teaching **this grade**:_____

Primary role: Teacher/co-teacher
Teacher Assistant

Total number of children in your class:_____

Sex: □ Male □ Female

Age: How old are you (in years)? _____

Race:

□ African American/Black □ White/Caucasian □ Asian/Pacific Islander
□ Native American/American Indian □ Other (please specify)_____

Are you of Hispanic origin? □ Yes □ No
If yes, what is your family's country of origin?_____

Please check the <u>highest</u> level of education that you have completed:

□ Less than High School Diploma or GED
□ High School Diploma or GED
□ Some College
□ Technical School
□ Bachelor's Degree
□ Master's Degree or Educational Specialist Degree
□ Doctoral Degree

Appendix C:
Student Identification Master List (to be Retained by the Teachers)
The University of Memphis

School: _____ Teacher 1: _____

Grade: _____ Teacher 2: _____

	Study ID Number
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Appendix D: Teacher Questionnaire Following Measurement

*Please complete this form **after** completing all three rating scales.*

	Not a Problem				Severe Problem
Overall, how would you rate this child's attention problems?	1	2	3	4	5
Overall, how would you rate this child's study skills?	1	2	3	4	5
Overall, how would you rate this child's overactivity?	1	2	3	4	5
Overall, how would you rate this child's social skills?	1	2	3	4	5
Overall, how would you rate this child's impulse control?	1	2	3	4	5

How long have you known this child? _____

How long has this child been in your class? _____

How old is this child? _____

What is the child's race? _____

I consulted with someone to complete this form: Yes No

If yes, with whom did you consult? _____

My teaching partner and I compared responses for this child: Yes No

Thank you for your time!

Appendix E

Description of Facets used in Analyses

Facet	*Description*
Classrooms (*c*)	Represents systematic differences across classrooms.
Students within Classrooms (*p:c*)	Represents differences in externalizing behavior symptoms across children.
Raters within Classrooms (*r:c*)	Represents systematic and overall differences in the way that raters within each classroom approach the task of completing the rating scale.
Instruments (*i*)	Represents a constant effect for students that stems from differences in item content across behavior rating scales.
Students by Raters within Classrooms (*pr:c*)	Represents inconsistencies of teachers' evaluation of particular children's behavior within each classroom.
Classrooms by Instruments (*ci*)	Represents inconsistencies across classrooms on ratings using the different instruments.
Error attributed to unspecified sources (*po*, *pi*, *ro*, *ri*, *proi*, *coi*, *e*)	Represents residual variance composed of the combination of *p*, *r*, *o*, and *i*, other facets that were not assessed but affect measurement, or random events.

Note. Reproduced from Bergeron et al., 2008. Reprinted with permission.

Appendix F

Institutional Review Board
Office of Sponsored Programs
University of Memphis
315 Admin Bldg
Memphis, TN 38152-3370

May 19, 2017

PI Name: Meredith Manguno
Co-Investigators:
Advisor and/or Co-PI:
Submission Type: Modification
Title: Dependability of the Ratings Across Time, Rater, and Form
IRB ID: #3980
Level of Review: Expedited

Approval: May 19, 2017
Expiration: *

*Modifications do not extend the expiration of the original approval

Approval of this project is given with the following obligations:

1. This IRB approval for modification has an expiration date, an approved renewal must be in effect to continue the project prior to that date. If approval is not obtained, the human consent form(s) and recruiting material(s) are no longer valid and any research activities involving human subjects must stop.

2. When the project is finished or terminated, a completion form must be submitted.

3. No change may be made in the approved protocol without prior board approval.

Thank you,
James P. Whelan, Ph.D.
Institutional Review Board Chair
The University of Memphis.