

# Sources of Bias in Teacher Ratings of Adolescents with ADHD

Brandon K. Schultz (Corresponding author)

Center for Intervention Research in Schools, Ohio University

200 Porter Hall, Athens, Ohio, 45701, USA

Tel: 1-740-597-3236 E-mail: schultb3@ohio.edu

Steven W. Evans

Center for Intervention Research in Schools, Ohio University

200 Porter Hall, Athens, Ohio, 45701, USA

Tel: 1-740-597-3236 E-mail: evanss3@ohio.edu

Received: January 6, 2012

Accepted: April 12, 2012

Published: May 1, 2012

doi:10.5539/jedp.v2n1p151

URL: <http://dx.doi.org/10.5539/jedp.v2n1p151>

*The research was financed by the Virginia Tobacco Settlement Fund (VTSF) in a grant awarded to the second author.*

## Abstract

Best practice assessment of childhood ADHD includes behavior ratings from multiple sources across multiple environments. However, adolescents in secondary schools interact with several teachers each day, and research has shown that teacher perceptions of the same child can be highly inconsistent. As a result, rating scale data can be equivocal, depending on which teachers are selected. The intent of the present study was two-fold: 1) to assess the consistency between teacher behavior ratings of adolescents with ADHD, and 2) to explore predictors of rater leniency or severity (i.e., sources of bias). Results suggest that interrater reliability within our sample was moderate, consistent with previous research. Further, teacher characteristics, including sex and age, were related to biases on ratings of student hyperactivity-impulsivity. Specifically, women and younger teachers provided significantly more severe ratings on average than did men and older teachers. Implications for the interpretation and statistical norming of ADHD rating scales are discussed.

**Keywords:** ADHD, Rating scales, Teacher ratings, Rater bias

## 1. Background

The cardinal symptoms of ADHD are significant and persistent impairment in attention or activity regulation relative to same-age peers. To be diagnosed with ADHD, individuals must exhibit six or more behavioral symptoms of inattention or hyperactivity-impulsivity for longer than six months, in two or more settings (e.g., home and school), with impairment in social, familial, or academic functioning beginning prior to age seven (American Psychiatric Association [APA], 2000). Thus, the diagnosis of ADHD is based entirely on observable behaviors and impairments, as research has failed to identify medical tests that reliably help in diagnosis (Pelham, Fabiano, & Massetti, 2005). Psychological tests also appear insufficient to diagnose ADHD because outcomes for individuals with and without the disorder largely overlap (i.e., poor instrument sensitivity and specificity), even though significant group differences are apparent between large samples (Frazier, Demaree, & Youngstrom, 2004). Similarly, group differences between ADHD and undiagnosed groups have been found on some neuropsychological measures, but these instruments are not sensitive or specific enough to diagnose individual cases (e.g., Homack & Riccio, 2004; Preston, Fennell, & Bussing, 2005). As a result, clinicians must rely primarily on behavioral observations—directly or indirectly—when assessing individual cases.

### 1.1 Interpreting Rating Scale Discrepancies

Behavior rating scale data from multiple informants have been generally shown to be both valid and useful and, as a result, behavior ratings are a vital component of best practice assessment of ADHD (American Academy of

Child and Adolescent Psychiatry, 1997; American Academy of Pediatrics, 2001). However, behavior ratings have significant limitations, including high rates of disagreement between raters. When behavior ratings are collected from multiple sources rating the same target child, some may appear relatively lenient and others appear relatively severe. Studies examining interrater reliability on behavior rating scales have traditionally found only moderate correlations (e.g., Achenbach, McConaughy, & Howell, 1987). In secondary schools, where students interact with several teachers in separate classrooms during the school day, interrater reliability among teachers (hereafter *inter-teacher reliability*) can be especially low. For example, Molina, Pelham, Blumenthal, and Galiszewski (1998) examined inter-teacher reliability among secondary teachers' ratings of adolescents with ADHD and found low to moderate reliability coefficients (intraclass correlations [ICCs] ranged from .21 to .52). Other studies suggest that inter-teacher reliability improves when teachers work within the same classroom, but even in overlapped environments there are considerable inter-teacher inconsistencies on ratings scales commonly used to assess ADHD (Danforth & DuPaul, 1996). So, although best practice recommendations suggest that clinicians collect data from multiple raters, the literature examining measurement reliability suggests that inter-teacher inconsistencies are common.

Of course, teachers have unique perspectives on student behavior, so differences between raters do not necessarily mean that some ratings are more valid or "true" than others. Rather, rating scales assess rater *perceptions*, which presumably offer an indirect measure of the target's actual behaviors (Smith, 2007). Thus, the term *bias* in this context does not necessarily refer to rater "error" or "prejudice", *per se*, but rather the general tendency for some raters' perceptions to appear relatively lenient and others to appear relatively severe. Moreover, rater bias is only a concern when ratings from multiple sources differ in meaningful ways (e.g., one rater indicates significant problems while another does not) and the goal of the assessment is to measure child behavior and *not* rater perception. So, in cases where teacher perception is the object of measurement, such as pre- and post-assessment of a classroom intervention, rater bias is generally not a concern. However, when a clinician attempts to assess a child's behaviors and impairments relative to a clinical standard or threshold, such as for diagnosis or service eligibility, rater bias can be troubling.

Conceptually, interrater inconsistencies are due to several factors, including changes in the target's behavior over time or contexts, interpersonal differences in rater perceptions, or random measurement error (e.g., rater misreads an item) (Kraemer, Measelle, Ablow, Essex, Boyce, & Kupfer, 2003). In their analysis of ADHD ratings, Gomez, Burns, Walsh, and Moura (2003) found that a significant proportion of variance was attributable to differences in rater perceptions (i.e., source effects) rather than actual changes in child behavior over time. For teachers, the variance associated with rater perception accounted for as much as 70% of the total variance in rating scale data. Gomez and colleagues arrived at this estimate using a confirmatory factor analytic method with a large sample of raters, which allowed for the mathematical estimation of variance components; but in actual practice, clinicians typically have no practical way of interpreting interrater discrepancies. For example, in most secondary school settings where teachers observe students in separate classrooms at separate times, it is impossible to determine whether discrepancies between teacher ratings reflect child behavior changes across contexts, differing teacher perceptions, or random error (DuPaul, 2003).

The issue of rater bias is rarely addressed in the professional literature. In an electronic search of the terms "rater+bias," "source+bias," or "teacher+bias" in articles available in the Psychology and Behavioral Sciences Collection database, only 37 unique results were found, spanning the years from 1966 to 2011. Similarly, rating scale manuals rarely discuss inter-teacher comparisons or methods for interpreting conflicting results (e.g., ADHD Rating Scale, Fourth Edition; DuPaul, Power, Anastopoulos, & Reid, 1998). As a result, clinicians are provided little guidance when confronted with inter-teacher discrepancies, outside of a few general strategies for combining or canceling discrepant data (e.g., Hart, Lahey, Loeber, & Hanson, 1994). Clinicians might sum the number of rating scale items that cross a predetermined threshold from *any* rater (e.g., Mitsis, McKay, Schulz, Newcorn, & Halperin, 2000) or sum only the threshold-level items endorsed by *multiple* raters (e.g., Power, Andrews, Eiraldi, Doherty, Ikeda, DuPaul, et al., 1998), but such strategies presume a qualitative parity among all raters and then summarize the results, rather than weigh interrater disagreement. In other words, the available strategies for handling rating scale discrepancies ignore potential sources of inter-teacher disagreement. Research is needed to systematically examine rater characteristics that predict relatively lenient or severe ratings of child behavior (Smith, 2007).

### 1.2 The Present Study

The present study examines inter-teacher discrepancies in ratings of adolescents with ADHD and offers a preliminary exploration of potential sources of these inconsistencies. We address two fundamental questions: First, how much agreement is there between teachers when rating adolescents with ADHD? Second, can

discrepancies between teacher ratings of the same adolescent be predicted? Although little research has examined our second question, some researchers hypothesize that rater idiosyncrasies unduly influence behavior ratings. For example, in their analysis of interrater reliability among teacher raters, Danforth and DuPaul (1996) conclude that "...characteristics of the teacher are a considerable source of error variance in ADHD rating scales" (p. 233). The second aim of the present study is an attempt to test this supposition.

For the first aim, we hypothesized that interrater reliability estimates would fall within the moderate range, roughly equivalent to those of previous research (e.g., Molina et al., 1998). For the second aim, specific hypotheses regarding potential sources of rater bias were based on related literatures due to the dearth of research specifically examining teacher bias. Specifically, we hypothesized that men would provide more lenient ratings than women, similar to trends found in broadband parent ratings of child psychopathology (e.g., Reynolds & Kamphaus, 2004). We also hypothesized that older teachers or teachers with relatively light workloads (e.g., small classes) would provide relatively lenient ratings when compared to younger teachers or teachers with relatively heavy workloads, based on research examining teacher burnout and professional development (e.g., Kokkinos, Panayiotou, & Davazoglou, 2004, 2005). Finally, we examined parental status based on our hypothesis that teachers make behavioral judgments of their students based partially on experiences with their own children. This last hypothesis is supported by research in related fields demonstrating that parental status affects perceptions of the causes and treatments for childhood health problems (e.g., obesity) (Hardus, van Vuuren, Crawford, & Worsley, 2003); but given the lack of parallel research specific to ADHD, a prediction as to the direction of this influence was untenable.

## 2. Method

The present study analyzed teacher rating scale data collected during the Challenging Horizons Program – Consultation Model study (CHP-C; Evans, Serpell, Schultz, & Pastor, 2007; Schultz, Evans, & Serpell, 2009), which was conducted in five rural middle schools in Virginia. The CHP-C study utilized a longitudinal, two-wave cohort design, with one cohort of student participants enrolled during their sixth through eighth grade years, followed immediately by a second cohort enrolled in the study in their sixth and seventh grade years.

### 2.1 Student Participants

In total, the CHP-C study enrolled 79 middle school students between the ages of 10 and 14 ( $Mdn = 11$  at intake). The majority of the CHP-C study sample was boys (77.2%), in a proportion roughly equivalent with the estimated sex ratio of children with ADHD in the general population (American Psychiatric Association [APA], 2000). Parents of the participants identified the majority of our sample (93.7%) as Caucasian, and most families (65.8%) reported a total yearly income less than \$60,001.

It was vital for the internal and external validity of the CHP-C study to establish that student participants met the diagnostic criteria for ADHD. To that end, clinical evaluations were conducted by trained graduate students under the supervision of a nationally certified school psychologist (first author) and licensed clinical psychologist (second author). All student participants accepted into the CHP-C study met the diagnostic criteria for one of the subtypes of ADHD, based on the criteria set forth by the Diagnostic and Statistical Manual – Fourth Edition – TR (DSM-IV-TR; APA, 2000). Twenty-eight participants (35.4%) met diagnostic criteria for Predominately Inattentive subtype, and the remaining 51 (64.6%) met diagnostic criteria for Combined subtype (inattention and hyperactivity/impulsivity). As part of the intake evaluation, student participants also completed the Kaufman Brief Intelligence Test (K-BIT; Kaufman & Kaufman, 1990) and the Wechsler Individual Achievement Test, Second Edition (WIAT-II; The Psychological Corporation, 2002) at intake. Based on these results, it appeared that our sample was of average intelligence (K-BIT FSIQ;  $M = 104.0$ ,  $SD = 11.8$ ), with average reading (WIAT-II Word Reading;  $M = 99.2$ ,  $SD = 13.4$ ), math (WIAT-II Numerical Operations;  $M = 94.2$ ,  $SD = 14.9$ ), spelling (WIAT-II Spelling;  $M = 96.9$ ,  $SD = 14.4$ ), and writing achievement (WIAT-II Written Expression;  $M = 96.9$ ,  $SD = 14.6$ ). For the purposes of the CHP-C study, candidates with estimated IQ scores below 80 or a comorbid diagnosis of post-traumatic stress disorder, substance use disorders, or psychosis were excluded because it was anticipated that the school-based interventions tested within the study (see Evans et al., 2007; Schultz et al., 2009) would be inappropriate for these populations.

### 2.2 Teacher Participants

We recruited teachers who participated in the second year of the CHP-C study ( $n = 108$ ), at the peak of student participant enrollment when both cohorts were active. To gather information about teachers, the researchers designed and administered a questionnaire (hereafter *Teacher Questionnaire*) that inquired about basic teacher demographic information and professional experiences. The questionnaire was sent to teachers at the beginning of the third year of the CHP-C study and included a cover letter that explained the purpose and design of the

present study.

Seventy-six teachers (70%) returned usable questionnaires. Of this group, 57 were women (75%). The median age of participating teachers was 43 years (range = 24 to 60), with an average of 15 years of teaching experience ( $SD = 10$ ). Most participating teachers reported that their highest degree was Bachelor's (70%), followed by Master's level training (25%), and then Master's with additional credits (5%). Teacher participants reported teaching an average of twenty-two students per class ( $SD = 4.8$ ), and an average of four classes per day ( $SD = 1.1$ ). More than two-thirds of teachers were parents (70%).

### 2.3 Procedures

All teachers received brief training on ADHD and aspects of the CHP-C study prior to the start of each school year. Much of the focus of these initial trainings was on explaining the protocol for completing and submitting the rating scales. No attempts were made to provide operational definitions of the individual rating scale items (e.g., defining "fidgetiness") or to build consensus on how specific behaviors should be rated. Rather, teachers read the identical rating scale instructions at the top of each form and interpreted rating scale items independently during each measurement occasion (Note 1).

The measurement procedures used to assess students at all five schools were identical. Brief rating scales (described in the next section) were administered to all core course teachers (reading, math, science, and social studies) on a monthly basis throughout the school year, with the exception of December due to the winter break. Teachers returned completed scales to a central location at their respective schools. At the end of the school year, all teachers were reimbursed for their participation at \$50 per child rated, which was considered adequate reimbursement to ensure high return rates without influencing the data.

In the present study, the researchers used teacher ratings from the spring semester of year two of the CHP-C study because this timeframe represented the peak enrollment of student participants. The overall return rate for teacher ratings during the targeted timeframe was 89%. Although the return rate was imperfect, this timeframe offered two advantages: First, students often do not exhibit academic impairments until the spring semester of each academic school year (Fallah, Buvinger, Evans, Schultz, & Serpell, 2006); consequently, teachers do not have sufficient opportunities to observe ADHD-related symptoms and impairments to make accurate assessments of their students in the fall semester. Second, previous research indicates that between-teacher reliability rates vary as a function of time, with large monthly fluctuations in the fall and stabilization of reliability rates in the spring. In one study it was found that the highest agreement rates were observed very early in the school year (September; Intraclass Correlation [ICC] = .70), but then coefficients quickly dropped to their lowest levels in November and December (ICC = .26 and .28, respectively), and then rebounded to levels between these two extremes (Evans, Allen, Moore, & Strauss, 2005). Thus, it appears that teacher ratings in the spring semester are likely to include an average amount of error variance when compared to overall yearly trends, possibly due to improved familiarity with the student over time.

### 2.4 Dependent Measures

#### 2.4.1 The Disruptive Behavior Disorder Scale (DBD; Pelham, Gnagy, Greenslade, & Milich, 1992)

The DBD is a narrow-band rating scale designed originally to measure the severity of symptoms associated with disruptive behavior disorders. The DBD was originally developed based on the DSM-III-R criteria for Attention Deficit Disorder (ADD), Oppositional Defiant Disorder (ODD), and Conduct Disorder (CD). To accurately capture these data, the DBD items are almost a verbatim inventory of the DSM-III-R behavioral symptom criteria. Research using teacher ratings of a random sample of boys suggested that the instrument had adequate internal consistency ( $\alpha = .96$ ). Further, individual items on the DBD appeared to have strong negative predictive power (NPP) for a full diagnosis of ADD (NPP rates per item all exceeded 0.95). In terms of positive predictive power (PPP), the items were not as strong (PPP rates ranged from 0.37 to 0.96), suggesting that items on the DBD were better at identifying students who did not meet the diagnostic criteria for Attention Deficit Disorder than for those who did (Pelham et al., 1992).

In the CHP-C, teachers were administered an updated version of the DBD scale, using virtually verbatim DSM-IV diagnostic criteria and shortened to include only the items relevant to ADHD. To our knowledge, only one study (Zuddas, Marzocchi, Oosterlaan, Cavolina, Ancilletta, & Sergeant, 2006) has examined the psychometrics of the updated DBD when used with teachers. In this study, DBD items measuring ADHD fit a two-factor structure (inattention and hyperactivity/impulsivity), consistent with the DSM-IV diagnosis, with very high rates of internal reliability within each subscale ( $\alpha \geq 0.93$ ).

#### 2.4.2 The Impairment Rating Scale (IRS; Fabiano et al., 2006)

The IRS is a broad-band scale that looks at several areas of impairment commonly associated with behavior disorders. The teacher version of the IRS consists of five items: two items relate to social functioning (with peers and with the teacher), two items relate to academics (academic progress and classroom functioning), and the last item asks the rater to consider if, overall, additional treatment or special services are required. Each item includes a line that the rater is asked to mark, with the left side of the line anchored by *No Problem, Definitely does not need treatment*, and the right side anchored by *Extreme Problem, Definitely needs treatment*. Items are scored by laying a transparent metric over top and then recording a score (0 to 6) based on the placement of the rater's mark.

According to preliminary studies of the IRS, the teacher version appears to have very high internal consistency ( $\alpha = .95$ ), adequate to very high test-retest reliability (Pearson  $r$ 's range from .74 to .96 with 3 to 4 month intervals between administrations), and good convergent and discriminant validity (Positive Predictive Power [PPP] = .90 and Negative Predictive Power [NPP] = .74). Studies suggest that while the items on the teacher version of the IRS appear to have less-than-adequate test-retest reliability (Median  $r = .56$ ), the overall interrater reliability between teachers and parents ( $r = .64$ ) is adequate, and appears comparable to that of other measures of impairment. In terms of convergent validity, the IRS appears to correlate strongly with the DBD ( $r$ s ranging from .67 to .85). Further, the IRS was found to have moderate to high correlations when compared to other teacher instruments that measure impairment (Fabiano et al., 2006).

#### 2.5 Statistical Analyses

Because the aims of this study were twofold, two separate statistical analyses were conducted. First, to assess the degree to which the teachers agreed in their ratings of ADHD symptoms and impairment, the agreement rates on the DBD and IRS were examined using intraclass correlations (ICCs). Due to an incomplete block measurement design (i.e., targets were nested within teacher teams and school sites), the analysis of agreement required the use of a one-way random effects model, consistent with Case 1 as described by Shrout and Fleiss (1979). In Case 1 ICC, the effects due to targets, raters, and the interaction effect between targets and raters are inseparable because targets are rated by various raters. Missing data resulted in varying numbers of teacher raters per target (i.e., unbalanced measurement design), so we corrected our ICCs by adjusting the degrees of freedom in the calculation of within-subjects variance (Bartko & Carpenter, 1976). In their study of between-teacher reliability, Molina and colleagues (1998) encountered a similar measurement scenario and addressed the complexities using identical procedures.

The second research question addressed by this study involved the degree to which teacher characteristics predicted rater bias. Based on the recommendations in the literature (e.g., Hill, O'Grady, & Price, 1988; Hoyt, 2002), a multiple regression analysis was used to assess the strength of potential predictors for the variance between raters. To quantify interrater disagreement, we subtracted each rating from the average rating per target per occasion, to derive an unstandardized deviation score. The unstandardized deviation scores were then used as the dependent variables in the regression analysis. To ensure independence of observations, we randomly selected one deviation score for each teacher participant from the target timeframe, so that no student-month pairings were repeated.

Five separate multiple regression analyses were conducted; the first three examined teacher responses to the three subscales of the DBD and the remaining two examined teacher responses on the academic and overall impairment items of the IRS. As described in the introduction, we examined four specific predictors: teacher sex, age, parental status, and workload. The predictors were entered into the analyses hierarchically in two blocks, beginning with ancillary teacher characteristics (sex and age), followed by a second block consisting of parental status and workload. The latter variable was computed by multiplying the average number of students per class by the number of classes taught per day, as reported on the Teacher Questionnaire.

### 3. Results

For the first aim of the study, we assessed consistency among teacher ratings using ICCs on the three subscales of the DBD and the academic and overall impairment items of the IRS. Descriptive statistics for these variables are provided in Table 1 and the ICCs for each of these variables are provided in Table 2. Overall, the ICCs ranged from 0.45 to 0.59, meaning that roughly 41% to 55% of the variance in teacher ratings was unexplained. These results are summarized in Table 2. The lowest ICCs on average were observed on the hyperactivity-impulsivity subscale of the DBD.

For the second aim of the study, we examined potential sources of rater bias by regressing unstandardized

deviation scores for teachers' ratings of target students onto teacher characteristics, including sex, age, parental status, and workload. To examine the relationship between these variables prior to analysis, we constructed a correlation matrix. The results are summarized in Table 3. Based on the zero-order correlations, it appears that the deviation scores on the rating scales were significantly and positively correlated with one another ( $ps < .01$ ), suggesting that rater leniency and severity trends were consistent across all dependent measures. Among the independent variables, it appears that teacher age and number of classes were significantly and positively correlated ( $r = .33, p < .01$ ), suggesting that older teachers taught more classes on average than younger teachers. Similarly, there appeared to be a significant and positive relationship between average number of students per class and number of classes taught per day ( $r = .32, p < .01$ ), suggesting that teachers who taught more classes per day reported larger class sizes than did teachers with fewer classes per day.

Next, we regressed unstandardized deviation scores from the inattention subscale of the DBD onto the indices of teacher characteristics and experiences in the regression model. For this and all subsequent regression analyses, we closely examined the model fit statistics, tolerance statistics, and the variance inflation factor to assess the likelihood of multicollinearity. We also examined casewise diagnostics to examine the likelihood of outliers, and visually scanned the standardized residuals plotted against the standardized predicted values to assess the likelihood of nonlinear relationships between the predictors and dependent variables. In all cases, these tests suggested that the statistical assumptions of multiple regression were met.

In the first analysis, which examined teacher ratings of student inattention, the saturated model did not explain a significant amount of the variance ( $F = 2.42, p = .06, \text{adjusted } R^2 = .07$ ). Thus, teacher characteristics did not predict a significant amount of teacher bias within ratings of student inattention. Next we examined teacher ratings of student hyperactivity-impulsivity. The saturated model appeared to explain a significant amount of the variance in teacher ratings ( $F = 2.85, p = .03, \text{adjusted } R^2 = .09$ ). Among the predictors in the full model, both teacher sex ( $\beta = .23, p = .048$ ) and teacher age ( $\beta = -.37, p = .040$ ) were statistically significant, suggesting that women and younger teachers provided more severe ratings of hyperactivity-impulsivity than men and older teachers, when the effects of the other predictors are held constant. These results are summarized in Table 4. In our final analysis of the DBD rating data, we also examined potential sources of rater bias on the total score. Again, the saturated model appeared to predict a significant proportion of rating variance ( $F = 3.35, p = .01, \text{adjusted } R^2 = .11$ ). Among the predictors in the model, teacher age ( $\beta = -.34, p = .02$ ), workload ( $\beta = .28, p = .02$ ), and teacher sex ( $\beta = .24, p = .02$ ) were statistically significant, suggesting that younger teachers, teachers with large workloads, and women provided more severe ratings than older teachers, teachers with small workloads, and men, when the effects of other predictors were held constant.

Next, we examined teacher ratings of academic impairment on the IRS. One teacher in the randomly selected sample failed to provide a response to this item, so this case was excluded from the analysis listwise, leaving a sample of 75 teachers rather than 76. The saturated model did not explain a significant proportion of the variance ( $F = 1.53, p = .20, \text{adjusted } R^2 = .03$ ), suggesting that disagreements among teacher ratings were unrelated to the examined predictors. Finally, we examined teacher ratings of overall impairment on the IRS and, again, the saturated model did not appear to explain a significant amount of the variance in teacher ratings ( $F = 2.36, p = .06, \text{adjusted } R^2 = .07$ ).

#### 4. Discussion

In this study we examined teacher behavior ratings of adolescents with ADHD. Specifically, we looked at consistency among teacher ratings and potential sources of bias within those ratings. On the first question, we found moderate rates of inter-teacher reliability, with intraclass correlations [ICCs] falling within the range of 0.45 to 0.59. All ICCs were statistically significant ( $ps < .001$ ), but suggest that a meaningful proportion of the variance—roughly 41% to 55%—was unexplained. These findings are consistent with previous research (e.g., Molina et al., 1998), suggesting that our sample of teachers provided a comparable amount of unexplained variance in their behavior and impairment ratings as other teacher samples providing data for adolescents with ADHD.

In our second analysis, we examined potential sources of rater bias. Specifically, we examined whether teacher sex, age, parental status, and workload predicted rater severity or leniency relative to ratings from other teachers. Across teacher ratings of ADHD-related behavior and impairment, our model was only successful at predicting rater trends on measures of hyperactivity-impulsivity and overall ADHD symptoms. Teacher characteristics did not appear to significantly influence ratings of ADHD-related impairments. On hyperactivity-impulsivity ratings, however, it appeared women teachers and younger teachers provided significantly more severe ratings than did men or older teachers when the effects of all predictors were held constant. Based on the unstandardized beta

values, our results suggest that women teachers provide more severe ratings with an average of 2.2 raw score points higher than men on the hyperactivity-impulsivity subscale of the DBD when the effects of other predictors were held constant. It also appeared that an increase of one standard deviation in teacher age (equaling almost 11 years) predicted .37 standard deviation, or 1.5 raw score points, greater leniency on the DBD hyperactivity-impulsivity subscale, when the effects of the other predictors were held constant. Similar results were found for overall teacher ratings of ADHD symptoms, but this finding is not surprising because hyperactivity-impulsivity is a component of that score. Readers should be cautious when interpreting this latter finding because ADHD rating scales have consistently been shown to measure two distinct constructs: inattention and hyperactivity-impulsivity (e.g., Zuddas et al., 2006). As a result, total scores do not offer unique information beyond the subscale scores. Still, it is interesting to note that, in addition to teacher sex and age, workload predicted rater bias on the total score of the DBD when all other predictors were held constant. For each increase of 35 students per day, total DBD raw scores increased by 2.1 points on average. In contrast to our initial hypotheses, teacher parental status did not predict a significant proportion of the variance in any of the observed ratings.

There are several potential interpretations of our findings. In regards to teacher age, older teachers within our sample may have had better classroom management skills than younger teachers. Research suggests that students with ADHD benefit from highly structured classrooms, and it seems particularly likely that, on average, teachers develop classroom management strategies over time and learn from previous experiences. Thus, the trend for relatively lenient ratings among the older teachers within our sample may be an artifact of students benefiting from well-structured classroom settings, which would conceivably impact student hyperactivity-impulsivity (e.g., talkativeness, disruptive behavior) more so than inattention (e.g., daydreaming, coming to class prepared). Or alternatively, older teachers may make judgments regarding student behavior relative to several cohorts of comparison students, even without improving classroom management skills. When compared to a large reference group of former students, some hyperactive behaviors may appear less extreme.

In regards to the effect of teacher sex, the relative leniency among men teachers on ratings of student hyperactivity-impulsivity might suggest that men observe the same behavior problems as women, but perceive these behaviors as less problematic. Similar research examining racial biases has found that teacher behavior ratings are influenced by target-rater similarities (e.g., Downey & Pribesh, 2004; Epstein, et al., 2005), providing some evidence for *dyad-specific* biases. Our sample was mostly boys, and it may be that men rate boys as less impaired than women due to identification with the target. Alternatively, adolescents with ADHD may respond differently to men and women teachers. A similar phenomenon appears to occur between mothers and fathers, as fathers are generally more lenient in their ratings than mothers (e.g., Reynolds & Kamphaus, 1992). Conceivably, this trend could be due to greater child obedience to fathers than mothers (Barkley, 2006, p. 97), which reduces the amount of behavior problems actually observed by fathers. It may hold that among teachers, these same sex differences exist, but the current results do not offer conclusive evidence for or against this hypothesis.

#### 4.1 Limitations

The present study has some noteworthy limitations, and readers should be cautious when interpreting the results. Perhaps the most limiting factor was the relatively small and demographically homogenous sample, as well as the naturalistic, incomplete block measurement design. In regard to our sample, teacher return rates on the Teacher Questionnaire limited our overall sample size. Prior to our regression analysis we calculated our statistical power based on the number of usable Teacher Questionnaires and found that with four predictors, our estimated power was .76 for each step in the hierarchical analysis, which is below the recommended level of .80 (Cohen, 1988). Thus, the present analysis was underpowered, thereby increasing the likelihood of a Type II error. Further, there was very little diversity within our teacher and student samples, so analyses of potential rater biases based on other factors (e.g., race) were unfeasible.

In regard to the measurement design, the imperfect return rates and complex relationships between teacher raters and students do not allow us to separate variance components associated with rater bias, true changes in target behavior, and other sources of error. The variance attributable to raters and the variance attributable to targets were partially nested and, from an analysis standpoint, completely confounded with measurement error (Brennan, 2001). As such, the results of our second aim (sources of rater bias) speak only to the predictive power of the teacher characteristics under examination. Still, if our predictors are reliable, our findings suggest that rater sex and age have implications for the interpretation of teacher ratings of ADHD symptoms.

#### 4.2 Implications

The results from the present study offer preliminary data to suggest that men and older teachers provide

relatively lenient ratings as compared to women and younger teachers when rating hyperactive-impulsive symptoms in adolescents. An implication of these findings, should they prove reliable, is that clinicians relying on teacher ratings to inform ADHD diagnosis or eligibility for services should consider characteristics of the raters when interpreting discrepant data from multiple informants. Meaningful variations in ratings are most likely to occur when differences between teacher raters are extreme. For example, based on our observed effect size, differences in ratings of student hyperactivity-impulsivity are likely to be meaningful once teacher age discrepancies go beyond 20 to 30 years of age. So, if a clinician has access to two teachers familiar with the target adolescent—a woman in the first year of her career and a man who has taught for 22 years—we would predict that, on average, the second teacher will provide relatively lenient ratings of hyperactivity-impulsivity. A cynical interpretation of this finding is that clinicians might manipulate assessment data to bring about desired results by privileging one rater over the other, or one data summarization strategy over another; but in our view, the lack of recognition of teacher bias in the research literature and in rating scale manuals already permits such manipulation.

Another implication of the present study is that clinicians should not presume parity among teacher raters, particularly when the purpose of assessment is to diagnose ADHD or to determine service eligibility relative to clinical thresholds. For example, if three teachers rate a student as having sub-threshold behavior symptoms and impairments whereas a fourth teacher's ratings suggest clinically significant hyperactivity-impulsivity, how does the clinician determine which result is accurate? Under the assumption of rater parity (i.e., the assumption that raters are comparable), clinicians might sum the rating scale items that cross the threshold for *any* rater, or sum only the threshold-level items endorsed by *multiple* raters. In this example, the first option would be tantamount to selecting the one clinically significant rating to represent all four teachers, potentially resulting in a Type I error. Alternatively, counting only the symptoms endorsed by multiple raters (i.e., symptom-wise cancellation) would dilute the one significant rating, potentially leading to a Type II error.

The present study offers preliminary data to suggest that ADHD rating scales are subject to significant rater biases based on teacher sex and age. Thus, cross-informant discrepancies in rating scale data appear to be partially predictable. As noted above, similar sex-related biases have been noted among parent rating scales, forcing separate norms for mothers and fathers to avoid misinterpretation (e.g., Reynolds & Kamphaus, 2004). Whether similar normative distinctions or other adjustments are needed for teacher rating scales of ADHD is unclear, but the results of the present study suggest that more research on teacher biases is warranted.

## References

- Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213-232. <http://dx.doi.org/10.1037//0033-2909.101.2.213>
- American Academy of Child and Adolescent Psychiatry (1997). Practice parameters for the assessment and treatment of children, adolescents, and adults with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, *36*, 85S-121S. <http://dx.doi.org/10.1097/00004583-199710001-00007>
- American Academy of Pediatrics (2000). Clinical practice guidelines: Diagnosis and evaluation of the child with attention-deficit/hyperactivity disorder. *Pediatrics*, *105*, 1158-1170. <http://dx.doi.org/10.1542/peds.105.5.1158>
- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., Text Revision). Washington, DC: Author.
- Barkley, R. A. (2006). *Attention-Deficit Hyperactivity Disorder* (3rd Ed.). New York: Guilford Press.
- Bartko, J. J., & Carpenter, W. T. (1976). On the methods and theory of reliability. *The Journal of Nervous and Mental Disease*, *163*, 307-317. <http://dx.doi.org/10.1097/00005053-197611000-00003>
- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Danforth, J. S., & DuPaul, G. J. (1996). Interrater reliability of teacher rating scales for children with Attention-Deficit Hyperactivity Disorder. *Journal of Psychopathology and Behavioral Assessment*, *18*, 227-237. <http://dx.doi.org/10.1007/BF02229046>
- Downey, D. B., & Pribesh, S. (2004). When race matters: Teachers' evaluations of students' classroom behaviors. *Sociology of Education*, *77*, 267-282. <http://dx.doi.org/10.1177/003804070407700401>

- DuPaul, G. J. (2003). Assessment of ADHD symptoms: Comment on Gomez et al. (2003). *Psychological Assessment, 15*, 115-117. <http://dx.doi.org/10.1037/1040-3590.15.1.115>
- DuPaul, G. J., Power, T. J., Anastopoulos, A. D., & Reid, R. (1998). *ADHD Rating Scale – IV: Checklists, Norms, and Clinical Interpretation*. New York: Guilford Press.
- Epstein, J. N., Willoughby, M., Valencia, E. Y., Tonev, S. T., Abikoff, H. B., Arnold, L. E., & Hinshaw, S. P. (2005). The role of children's ethnicity in the relationship between teacher ratings of attention-deficit/hyperactivity disorder and observed classroom behavior. *Journal of Consulting and Clinical Psychology, 73*, 424-434. <http://dx.doi.org/10.1037/0022-006X.73.3.424>
- Evans, S. W., Allen, J. Moore, S., & Strauss, V. (2005). Measuring symptoms and functioning of youth with ADHD in middle schools. *Journal of Abnormal Child Psychology, 33*, 695-706. <http://dx.doi.org/10.1007/s10802-005-7648-0>
- Evans, S. W., Serpell, Z. N., Schultz, B. & Pastor, D. (2007). Cumulative benefits of secondary school-based treatment of students with ADHD. *School Psychology Review, 36*, 256-273. [Online] Available: <http://www.nasponline.org/publications/spr/index.aspx?vol=36&issue=2>
- Fabiano, G. A., Pelham, W. E., Waschbusch, D. A., Gnagy, E. M., Lahey, B. B., Chronis, A. M., ... Burrows-Maclean, L. (2006). A practical measure of impairment: Psychometric properties of the impairment rating scale in samples of children with attention deficit hyperactivity disorder and two school-based samples. *Journal of Clinical Child and Adolescent Psychology, 35*, 369-385. [http://dx.doi.org/10.1207/s15374424jccp3503\\_3](http://dx.doi.org/10.1207/s15374424jccp3503_3)
- Fallah, N., Buvinger, E., Evans, S. W., Schultz, B., & Serpell, Z. (2006, August). *Outcomes of a consultation model school-based psychosocial intervention for middle School Students with ADHD*. Poster presented at the Annual Meeting of the American Psychological Association, New Orleans, LA.
- Frazier, T. W., Demaree, H. A., & Youngstrom, E. A. (2004). Meta-analysis of intellectual and neuropsychological test performance in attention-deficit/hyperactivity disorder. *Neuropsychology, 18*, 543-555. <http://dx.doi.org/10.1037/0894-4105.18.3.543>
- Gomez, R., Burns, G. L., Walsh, J. A., & de Moura, M. A. (2003). A multitrait-multisource confirmatory factor analytic approach to the construct validity of ADHD rating scales. *Psychological Assessment, 15*, 3-16. <http://dx.doi.org/10.1037/1040-3590.15.1.3>
- Hardus, P. M., van Vuuren, C. L., Crawford, D., & Worsley, A. (2003). Public perceptions of the causes and prevention of obesity among primary school children. *International Journal of Obesity, 27*, 1465-1471. <http://dx.doi.org/10.1038/sj.ijo.0802463>
- Hart, E. L., Lahey, B. B., Loeber, R., & Hanson, K. S. (1994). Criterion validity of informants in the diagnosis of disruptive behavior disorders in children: A preliminary study. *Journal of Consulting and Clinical Psychology, 62*, 410-414. <http://dx.doi.org/10.1037//0022-006X.62.2.410>
- Hill, C. E., O'Grady, K. E., & Price, P. (1988). A method for investigating sources of rater bias. *Journal of Counseling Psychology, 35*, 346-350. <http://dx.doi.org/10.1037//0022-0167.35.3.346>
- Homack, S., & Riccio, C. A. (2004). A meta-analysis of the specificity and sensitivity of the Stroop Color and Word Test for children. *Archives of Clinical Neuropsychology, 19*, 725-743. <http://dx.doi.org/10.1016/j.acn.2003.09.003>
- Hoyt, W. T. (2002). Bias in participant ratings of psychotherapy process: An initial generalizability study. *Journal of Counseling Psychology, 49*, 35-46. <http://dx.doi.org/10.1037//0022-0167.49.1.35>
- Kaufman, A. S., & Kaufman, N. L. (1990). *Kaufman Brief Intelligence Test Manual*. Circle Pines, MN: American Guidance Service.
- Kokkinos, C. M., Panayiotou, G., & Davazoglou, A. M. (2005). Correlates of teacher appraisals of student behaviors. *Psychology in the Schools, 42*, 79-89. <http://dx.doi.org/10.1002/pits.20031>
- Kokkinos, C. M., Panayiotou, G., & Davazoglou, A. M. (2004). Perceived seriousness of pupils' undesirable behaviours: The student teacher's perspective. *Educational Psychology, 24*, 109-120. <http://dx.doi.org/10.1080/0144341032000146458>
- Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry, 160*, 1566-1577.

<http://dx.doi.org/10.1176/appi.ajp.160.9.1566>

Mitsis, E. M., McKay, K. E., Schulz, K. P., Newcorn, J. H., & Halperin, J. M. (2000). Parent-teacher concordance for DSM-IV attention-deficit disorder in a clinic-referred sample. *Journal of the American Academy of Child and Adolescent Psychiatry*, *39*, 308-313. <http://dx.doi.org/10.1097/00004583-200003000-00012>

Molina, B., Pelham, W. E., Blumenthal, J., & Galiszewski, E. (1998). Agreement among teachers' behavior ratings of adolescents with a childhood history of attention deficit hyperactivity disorder. *Journal of Clinical Child Psychology*, *27*, 330-339. [http://dx.doi.org/10.1207/s15374424jccp2703\\_9](http://dx.doi.org/10.1207/s15374424jccp2703_9)

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 and Adolescent Psychology*, *34*, 449-476. [http://dx.doi.org/10.1207/s15374424jccp3403\\_5](http://dx.doi.org/10.1207/s15374424jccp3403_5)

Pelham, W. E., Gnagy, E. M., Greenslade, K. E., & Milich, R. (1992). Teacher ratings of DSM-III symptoms for the disruptive behavior disorders. *Journal of the American Academy of Child and Adolescent Psychiatry*, *31*, 210-218. <http://dx.doi.org/10.1097/00004583-199203000-00006>

Power, T. J., Andrews, T. J., Eiraldi, R. B., Doherty, B. J., Ikeda, M. J., DuPaul, G. J., & Landau, S. (1998). Evaluating attention deficit hyperactivity disorder using multiple informants: The incremental utility of combining teacher with parent reports. *Psychological Assessment*, *10*, 250-260. <http://dx.doi.org/10.1037//1040-3590.10.3.250>

Preston, A. S., Fennell, E. B., & Bussing, R. (2005). Utility of a CPT in diagnosing ADHD among a representative sample of high risk children: A cautionary study. *Child Neuropsychology*, *11*, 459-469. <http://dx.doi.org/10.1080/09297040591001067>

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

Schultz, B. K., Evans, S. W., & Serpell, Z. N. (2009). Preventing failure among middle school students with attention deficit hyperactivity disorder: A survival analysis. *School Psychology Review*, *38*, 14-27. [Online] Available: <http://www.nasponline.org/publications/spr/index.aspx?vol=38&issue=1>

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420-428. <http://dx.doi.org/10.1037//0033-2909.86.2.420>

Simonoff, E., Pickles, A., Hewitt, J., Silberg, J., Rutter, M., Loeber, R., ... Eaves, L. (1995). Multiple raters of disruptive child behavior: Using a genetic strategy to examine shared views and bias. *Behavior Genetics*, *25*, 311-326. <http://dx.doi.org/10.1007/BF02197280>

Smith, S. R. (2007). Making sense of multiple informants in child and adolescent psychopathology: A guide for clinicians. *Journal of Psychoeducational Assessment*, *25*, 139-149. <http://dx.doi.org/10.1177/0734282906296233>

The Psychological Corporation. (2002). *Wechsler Individual Achievement Test – Second Edition: Examiner's Manual*. San Antonio: Harcourt Brace & Company.

Zuddas, A., Marzocchi, G. M., Oosterlaan, J., Cavolina, P., Ancilletta, B., & Sergeant, J. (2006). Factor structure and cultural factors of disruptive behavior disorders symptoms in Italian children. *European Psychiatry*, *21*, 410-418. <http://dx.doi.org/10.1016/j.eurpsy.2005.08.001>

## Note

Note 1. A subset of student participants ( $n = 43$ ) received a school consultation treatment as part of the CHP-C study (see Evans et al., 2007; Schultz et al., 2009), but randomization was conducted by school, thereby ensuring that no teacher rated students from both conditions. Rather, teacher teams in the present study always reported on targets in the same experimental condition, thereby making comparisons between teachers (e.g., leniency and severity) meaningful, regardless of the influence of the CHP-C interventions.

Table 1. Descriptive statistics for teacher ratings on the DBD and IRS

	<i>M</i>	<i>SD</i>	Min	Max	Skew	Kurt
<b>DBD Subscales</b>						
Inattention <sup>a</sup>	13.0	7.7	0	27	0.1	-1.0
Hyperactivity-Impulsivity <sup>a</sup>	7.8	6.9	0	27	0.8	-0.1
Total Score <sup>b</sup>	20.7	13.4	0	54	0.4	-0.6
<b>IRS Items</b>						
Academic Impairment <sup>c</sup>	3.1	2.2	0	6	-0.1	-1.4
Overall Impairment <sup>c</sup>	2.9	2.2	0	6	0.0	-1.5

Note. DBD = Disruptive Behavior Disorders scale; IRS = Impairment Rating Scale.

<sup>a</sup>Subscales of the DBD range from 0 to 27.

<sup>b</sup>Total score of the DBD ranges from 0 to 54.

<sup>c</sup>Items of the IRS range from 0 to 6.

Table 2. Intraclass Correlations for Teacher Ratings of ADHD Symptoms and Impairment by Month

	Feb ( <i>n</i> = 62)	March ( <i>n</i> = 58)	April ( <i>n</i> = 66)	May ( <i>n</i> = 60)
<b>DBD Subscales</b>				
Inattention	0.52	0.55	0.55	0.55
Hyperactivity-Impulsivity	0.48	0.46	0.48	0.52
Total Score	0.51	0.53	0.55	0.56
<b>IRS Items</b>				
Academic Impairment	0.52	0.58	0.56	0.59
Overall Impairment	0.50	0.45	0.53	0.50

Note. DBD = Disruptive Behavior Disorders scale; IRS = Impairment Rating Scale. All correlations were statistically significant ( $p < .001$ ).

Table 3. Correlation matrix for the independent (predictor) and dependent variables

	S	TA	PS	NS	NC	DV1	DV2	DV3	DV4	DV5
Sex (S) <sup>a</sup>	1.00									
Teacher Age (TA)	.20	1.00								
Parental Status (PS) <sup>b</sup>	.22	** .56	1.00							
Number of Students (NS)	-.09	.12	.02	1.00						
Number of Classes (NC)	.02	** .33	.11	** .32	1.00					
Dev. on DBD-IA (DV1)	.17	.00	.14	.19	.16	1.00				
Dev. on DBD-HI (DV2)	.19	-.15	.07	-.02	.20	** .53	1.00			
Dev. on DBD Total (DV3)	.21	-.08	.12	.10	.21	** .88	** .87	1.00		
Dev. on IRS Acad. (DV3)	.15	.05	.15	* .24	.12	** .66	** .40	** .61	1.00	
Dev. on IRS Total (DV4)	.14	-.02	.12	.20	.20	** .63	** .51	** .65	** .85	1.00

Note. Dev. = Deviation Score; DBD-IA = Inattention Subscale of DBD; DBD-HI = Hyperactivity-Impulsivity Subscale of DBD; IRS Acad. = Academic Item on IRS

<sup>a</sup>Man = 1, Woman = 2. <sup>b</sup>No = 0, Yes = 1.

\*  $p < .05$  (two-tailed) \*\*  $p < .01$  (two-tailed)

Table 4. Hierarchical multiple regression results for the analysis of bias in teacher ratings of inattention and hyperactivity-impulsivity symptoms

	DBD-HI <sup>a</sup>			DBD-IA <sup>b</sup>		
	$\Delta R^2$	<i>B</i>	<i>SE B</i> $\beta$	$\Delta R^2$	<i>B</i>	<i>SE B</i> $\beta$
Step One	.07			.03		
Constant		-0.53	2.41		-2.16	2.58
Sex		2.16	1.08 .23		1.78	1.16 .18
Age		-0.07	0.04 -.19		-0.01	0.05 -.04
Step Two	.07			.09*		
Constant		-1.19	2.55		-3.37	2.69
Sex		2.15	1.07 .23*		1.85	1.13 .19
Age		-0.14	0.05 -.37*		-0.09	0.06 -.22
Parental Status		1.83	1.19 .21		1.75	1.25 .19
Workload		0.03	0.01 .21		0.03	0.01 .27

Note. DBD-HI = Teacher hyperactivity-impulsivity ratings on the Disruptive Behavior Disorders rating scale; DBD-IA = Teacher inattention ratings on the Disruptive Behavior Disorders rating scale.

<sup>a</sup> Model  $F = 2.85, p = .03$ . <sup>b</sup> Model  $F = 2.42, p = .06$ .

\*  $p < .05$