HHS Public Access

Author manuscript

J Clin Child Psychol. Author manuscript; available in PMC 2016 May 18.

Published in final edited form as:

J Clin Child Psychol. 1998 October ; 27(3): 330–339. doi:10.1207/s15374424jccp2703_9.

Agreement Among Teachers' Behavior Ratings of Adolescents With a Childhood History of Attention Deficit Hyperactivity Disorder

Brooke S. G. Molina,
Western Psychiatric Institute and Clinic, University of Pittsburgh School of Medicine

William E. Pelham,
State University of New York at Buffalo

Jonathan Blumenthal, and
Western Psychiatric Institute and Clinic, University of Pittsburgh School of Medicine

Emily Galiszewski
Western Psychiatric Institute and Clinic, University of Pittsburgh School of Medicine

Abstract

Examined agreement among secondary school teachers' behavior ratings for 66 adolescent boys with a history of attention deficit hyperactivity disorder. Behavior ratings consisted of the Teacher Report Form, Iowa/Abbreviated Conners, and the Disruptive Behavior Disorders Rating Scale. Ratings from 2 to 5 teachers were collected for each adolescent. In contrast to previous studies, agreement was examined using statistical indices that corrected for chance agreement and discrepancies in scores (i.e., intraclass correlation [ICC], kappa) in addition to traditional indices (i.e., Pearson correlation and percentage agreement) typically used in the relatively sparse literature on teacher agreement for adolescent behavior ratings. Agreement was poor for dimensional subscale scores (Pearson correlations were in the .40–.50 range, and ICCs were in the .20–.50 range) as well as for categorization of youth as above or below clinical cutoffs (percentage agreement was between 52% and 96%, but ICCs and kappas ranged from .17 to .57). Findings suggest that, regardless of behavior rating scale used, a multiple teacher assessment strategy should be adopted for clinical assessment, treatment design, and evaluation of treatment efficacy.

Teacher ratings of child behavior are tremendously useful when a goal of clinical work or research is to assess the disruptive behavior of a child (i.e., overactivity, impulsivity, attentional difficulties, defiance, or severe conduct problems). Not only is determination of treatment efficacy dependent upon teachers' observations, but adequate diagnosis of disruptive behavior disorders depends upon their input. For example, diagnosis of attention deficit hyperactivity disorder (ADHD) requires determination of impaired functioning in

Requests for reprints should be sent to Brooke Molina, Western Psychiatric Institute and Clinic, University of Pittsburgh School of Medicine, 3811 O'Hara Street, Pittsburgh, PA 15213. molinab@msx.umpc.edu.
Jonathan Blumenthal is now with the National Institute of Mental Health, Bethesda, MD.

Portions of this study were presented at the 1996 meeting of the American Psychological Association, Toronto, Ontario, Canada.

more than one domain (*Diagnostic and Statistical Manual of Mental Disorders*, 4th ed. [*DSM–IV*]; American Psychiatric Association, 1994). Because children spend so much of their time in school, functioning in the academic domain is of particular importance. Furthermore, academic performance is often severely impaired by difficulties with behaviors such as attentional problems and noncompliance, making teachers especially salient and ecologically valid reporters of child functioning. Research has also shown that teacher ratings are more sensitive than parent ratings to stimulant medication effects (Sprague, Christensen, & Werry, 1974), which makes collection of teacher reports imperative when a goal of treatment is to determine medication efficacy. Clearly, examination of a child's functioning in school should always be included in comprehensive assessments and treatment of childhood disruptive behavior disorders.

An increasingly popular means of collecting teacher observations is to use behavior rating scales (Hutton, Dubes, & Muir, 1992). These measures are typically characterized by (a) standard instructions and response formats, (b) multiple items for assessing competencies and problems, (c) ability to sum individual items to produce indices of functioning in specific areas, (d) normative samples, and (e) reliability and validity data (McConaughy, 1993). As a result, because of their ease of administration, behavioral focus, and capacity for facilitating communication among parents, teachers, and other professionals, certain rating scales have become quite popular (e.g., the Teacher Report Form [TRF]; Achenbach & Edelbrock, 1986; the Conners Revised Teacher Rating Scale; Goyette, Conners, & Ulrich, 1978). However, their use for children has been developed over a longer period of time than for adolescents. In particular, one issue that has received relatively little attention is cross-informant agreement when adolescents are being evaluated. More specifically, there is little information available regarding the extent to which behavior ratings by secondary school teachers converge. The interrater agreement of secondary school teachers has relevance for both clinical and research assessment strategy as well as for under-standing the heterogeneity of expression of adolescent behavior disorders.

Achenbach and colleagues have addressed this issue indirectly in their well-known meta-analysis of cross-informant correlations (Achenbach, McConayghy, & Howell, 1987). They determined that the average correlation between teachers' ratings of the same child was .64; which suggested a moderately high degree of interobserver consistency. They concluded that data from a single informant, where other informants would see the child under generally similar conditions (e.g., school), would typically be adequate. However, in one of his guides to use of the TRF, Achenbach recommends obtaining "TRFs from whichever teachers know the child reasonably well" (Achenbach, 1991, p. 109). It appears that most investigators of child behavior problems, at least in empirical studies, have chosen procedures consistent with the former recommendation. Recent studies making use of teacher behavior ratings of adolescents typically report that only one teacher's rating was used (e.g., Greenbaum, Dedrick, Prange, & Friedman, 1994; Lee, Elliott, & Barbour, 1994; Phares, Compas, & Howell, 1989), and there is usually little information about which teacher was chosen for this purpose. A careful examination of the Achenbach et al. (1987) review, however, reveals several reasons why this strategy may not be adequate. First, of 20 teacher-teacher correlations examined by Achenbach et al., only 3 were derived from samples of adolescents, suggesting that the estimate of .64 was appropriate for primary rather than for

secondary school children. Second, some of the teacher-teacher correlations were obtained from ratings made by teachers and their aides working in the same classroom, which would yield higher agreement than correlations obtained from ratings made by secondary school teachers from different classrooms (the more likely occurrence in middle schools and high schools). Third, when correlations aggregated across different types of informants (e.g., parents, teachers, child self-report) were examined separately for children and adolescents, agreement was lower for adolescents ($r = .41$) than for children ($r = .51$). Considered together, these findings suggest that agreement among secondary school teachers is probably low (or at least lower than .64) and that ratings from more than one teacher may need to be collected for an accurate picture of functioning.

With the exception of one published study, there is little empirical data addressing this issue directly. Simpson (1991) examined agreement among high school teachers' ratings of students using the Revised Behavior Problem Checklist (Quay & Petersen, 1983, 1984), an instrument widely used as a screen for behavioral disorders. Correlations between teachers (two for each student) ranged from −.09 to .54 across the different subscales, with an average correlation of .25. When scores were recoded to reflect normal, mildly deviant, or highly deviant functioning, overlap between teachers' ratings was minimal. Even more importantly, the highest level of agreement (approximately 30%), which was still quite low, was for behavior problems that are readily observed by virtue of their overt expression (conduct disorder [CD], attention problems-immaturity, and motor tension-excess). Overlap among teacher ratings was substantially lower for other problems, such as anxiety-withdrawal. Thus, even when teachers are rating behaviors that are quite visible and easy to recognize, there is considerable variability in their perceptions of behavioral difficulty.

There are good reasons to expect that agreement among teachers' ratings of adolescents should be low, based on developmental and contextual changes occurring for children in the adolescent years. First, although preadolescent children typically have one classroom teacher, adolescents in secondary school can have as many as eight teachers in a single day. The impact of this change can be far-reaching for a child with disruptive behavior problems. For example, children who have difficulty regulating attention and impulse control may respond inconsistently across classrooms with different levels of environmental structure and teacher tolerance, creating frustration for pupil and faculty alike. Variability in teacher skills and student aptitude across subjects may further affect expression of behavior. It is unclear, however, the extent to which these variations affect teacher ratings—the very index frequently relied on as objective assessment. For instance, just how much variability across teachers' ratings can be expected, and is this variability affected by the choice of instrument? These questions were the focus of the current study.

A key contribution of the current study was our method of examining agreement. Previous studies of agreement among informant reports of child behavior have relied on Pearson correlations and percentage agreement indices (Achenbach et al., 1987). Although the Pearson correlation provides an index of association between pairs of raters, it does not index actual disagreement in rating levels (Bartko & Carpenter, 1976). Percentage agreement, while having computational and intuitive appeal, ignores chance agreement, which can be plentiful when few categories are used by raters (Bartko & Carpenter, 1976;

Hartmann, 1977; Spitzer, Cohen, Fleiss, & Endicott, 1967). Consequently, previous studies of interobserver agreement may have overestimated the extent to which reports converge. For this study, in addition to providing the traditional measures of agreement (Pearson correlations and percentage agreement) for comparison purposes, intraclass correlations and kappas are provided as indices of association that correct for these limitations (Bartko & Carpenter, 1976; Hartmann, 1977; Shrout & Fleiss, 1979; Spitzer et al., 1967).

Method

Participants

Participants were 88 adolescent boys with a history of treatment for ADHD in a summer day-treatment program for ADHD at Western Psychiatric Institute and Clinic, University of Pittsburgh Medical Center. Adolescents and their parents were participating in a larger study of teenagers with a history of ADHD. Participants were selected for follow-up interviews if they had received a *Diagnostic and Statistical Manual of Mental Disorders*, third edition, revised (*DSM-III-R*; American Psychiatric Association, 1987) or *DSM-IV* diagnosis of ADHD at the time of initial evaluation for the summer program (the earliest evaluations were in 1987 and the most recent were in 1994). Diagnoses were based on structured clinical interviews with parents by master's- and doctoral-level clinicians and by teacher report using the Disruptive Behavior Disorders Scale (DBD; Pelham, Gnagy, Greenslade, & Milich, 1992). Exclusionary criteria for follow-up included IQ less than 80, history of seizures or other neurological problems, history of pervasive developmental disorder, schizophrenia, or other psychotic disorders, sexual disorders, or organic mental disorders.

At follow-up, adolescents ranged in age from 13 to 18 ($M = 15.15$, $SD = 1.43$). Most were White (91% self-reported as Caucasian/White, 7% as African American/Black, and 2% as other). They were from a wide range of socioeconomic backgrounds (parent education ranged from partial high school to graduate professional, and total annual household income ranged from $10,000 to $300,000). The median level of parent education (same for mothers and fathers) was partial college, and the median total annual household income was $45,000. Most adolescents attended public secondary schools (73%), 13% attended private secondary schools, 3% attended vocational training institutes, and 11% attended school in other specialized settings (e.g., partial day-treatment facility, correctional institute). Between 18% and 30% had a learning disability, which is consistent with studies of ADHD children (e.g., Barkley, 1990) and with reports that adolescents with ADHD (or childhood histories of ADHD) have academic difficulties (August & Garfinkel, 1990; Barkley, Fischer, Edelbrock, & Smallish, 1990).[1]

Procedure

After collecting written informed consent from both parents and adolescents, a packet of questionnaires was sent to each adolescent's guidance counselor. Each packet contained, in

[1]Because there is not one commonly agreed-upon definition of learning disability (LD), we calculated the percentage of learning disabled adolescents in our sample using two procedures described by Barkley (1990). Thirty percent met criteria for LD as defined by a 15-point discrepancy between IQ and either math, reading, or spelling standard scores ($M = 100$, $SD = 15$). Eighteen percent met criteria for LD after meeting the 15-point discrepancy criterion, but they also had math, reading, or spelling achievement scores that were 1.5 standard deviations below the mean (a score of 100).

addition to a request for other information (i.e., grades, attendance, achievement test scores, and school schedule), five sets of three questionnaires each: the Achenbach TRF (Achenbach & Edelbrock, 1986), the Iowa/Abbreviated Conners Teacher Rating Scale (IOWA; Goyette et al., 1978; Loney & Milich, 1982), and DBD (Pelham, Gnagy, et al., 1992). Each adolescent's guidance counselor was instructed to distribute questionnaires to primary academic course teachers (e.g., English, social studies, math, etc.).[2] The number of questionnaires returned is displayed in Table 1. An average of four teachers per adolescent returned questionnaires, but the number of questionnaires returned ranged from a low of zero (for 11 adolescents, no questionnaire data were available for reasons such as parents refused consent or the school failed to return the questionnaires) to a high of five (except in one instance where eight sets of questionnaires were returned). This left 66 adolescents for whom two or more teachers provided ratings. Because most analyses were based on the 66 adolescents with ratings from two or more teachers, comparisons between these youths and the remaining 22 were made on demographic and disruptive behavior variables to determine extent of sampling bias. There were no statistically significant differences between groups on ethnicity, adolescent age, parent education, family income, and number of disruptive behavior symptoms on the DBD and IOWA. Of the 258 teachers who completed questionnaires, 78% were regular education teachers and 22% were special education teachers. Sixty-eight percent of the teachers who returned questionnaires taught primary academic courses (i.e., math, social studies, English, and science), and 32% taught other courses such as art, music, or gym.

Measures—The TRF (Achenhach & Edelbrock, 1986) is widely used by clinicians, teachers, and researchers and it has well-established reliability and validity (Achenbach, 1991). Factor analyses with the TRF have found three separate disruptive behavior factors relevant to this study which have well-established psychometric properties. Subscale scores for these factors (Attention Problems, Delinquent Behavior, and Aggressive Behavior) were used for analyses involving dimensional variables. However, categorical discrimination between presence and absence of problems was also coded using *T* score cutoffs at the bottom of the clinical range for each syndrome (Achenbach, 1991). That is, dimensional subscale scores above the cutoff resulted in a categorical score of one, whereas dimensional subscale scores below the cutoff resulted in a categoripal score of zero. The TRF also asks teachers how Well they know the students they are rating. Most teachers (66%) reported knowing students *moderately well* on a scale from 1 (*not well*) to 3 (*very well*), suggesting that guidance counselors may have chosen teachers who knew their students well enough to rate them knowledgeably.

The IOWA is also a commonly used screening instrument for behavior problems. Both the inattention/overactivity (IO) and oppositional defiant (OD; sometimes referred to as aggression) factors, identified in previous research have made significant and unique contributions toward predicting observed classroom behaviors (Atkins, Pelham, & Licht, 1989). This brief measure consists of 15 items that assess difficulties with attention, overactivity, impulsivity, defiance, and moodiness. Response options for each item ranged

[2]Teachers providing ratings at follow-up in adolescence were different from teachers providing ratings for entry into the summer treatment program.

from 0 (*not at all*) to 3 (*very much*). Dimensional subscale scores were calculated for the IO and OD factors, and research cutoff scores (Pelham, Milich, Murphy, & Murphy, 1989) were used to identify adolescents functioning in the clinically impaired range.[3] Estimates of internal consistency were in the acceptable range (IO, .79; OD, .87).

To obtain information necessary for making *DSM-IV* diagnoses, the DBD (Pelham, Gnagy, et al., 1992) was used to assess teacher endorsement of symptoms for ADHD, oppositional defiant disorder (ODD), and CD. This measure, which reflects the symptoms listed in the *Diagnostic and Statistical Manual of Mental Disorders*, third edition (*DSM–III*; American Psychiatric Association, 1980), *DSM–III–R*, and *DSM–IV* for these disorders, consists of 45 items each with four close-ended response options ranging from *not at all* to *very much*. The numbers of symptoms endorsed as occurring *pretty much* or *very much* within each of the previously established DBD factors (inattention, impulsivity-overactivity, OD, and CD; Pelham, Evans, Gnagy, & Greenslade, 1992; Pelham, Gnagy, et al., 1992; Pillow, Pelham, Hoza, Molina, & Stultz, 1998) were used to determine diagnostic status. These categorical variables reflected the presence or absence of ADHD (Inattentive Type, Hyperactive/Impulsive Type, or Combined Type), ODD, or CD. Dimensional subscale scores were also calculated as the mean of the respective DBD factor items (inattention, 9 items; impulsivity-overactivity, 10 items; OD, 13 items; and CD, 13 items). Estimates of internal consistency for the dimensional subscale scores are as follows: inattention, .67; impulsivity-overactivity, .67; OD, .81; and CD, .92.

Results

Agreement for Dimensional Subscale Scores

Intraclass correlations between dimensional sub-scale scores obtained from different teachers are presented in the first column of Table 2. Each intraclass correlation reflects agreement between ratings by all teachers for each adolescent (i.e., of the 66 adolescents for whom two or more teachers completed ratings).[4] The average pairwise Pearson correlation for each subscale is listed in the second column of Table 2 (Pearson correlations were temporarily transformed to Fisher z values before weighting by the number of teachers in each pair and averaging). Finally, for direct comparison to previous studies, Table 2 (third column) also shows the Pearson *r* correlations between ratings made by two teachers who were randomly selected for each adolescent. To control for Type 1 error, correlations are marked as statistically significant at $p < .01$ or better.

The intraclass correlations in Table 2 show that, with the exception of DBD CD (for which the correlation is quite low), agreement ranged from a low of .21 for the IOWA OD scale to

[3]In the absence of IOWA score norms for adolescents, cutoff scores were chosen from a sample of fourth- and fifth-grade children (Pelham et al., 1989). These cutoff scores may be conservative for adolescents. However, we are comforted by the findings in Table 3 showing either similarity across measures in percentage of cases diagnosed (e.g., 43% of adolescents rated by one or more teachers as having attention problems using the IOWA vs. 45% of adolescents rated as such with the DBD) or higher estimates with the IOWA (e.g., 33% rated as having oppositional problems with the IOWA vs. 18% rated as such with the DBD). Nevertheless, we recommend caution when interpreting the IOWA findings due to our use of childhood norms.

[4]Several intraclass correlations are available for computation using Statistical Package for the Social Sciences macros written by David Nichols, Senior Support Statistician (accessible via the World Wide Web at http://www.spss.com). We used Case 1 from the Shrout and Fleiss (1979) article, with the correction for varying numbers of raters per participant listed in Bartko and Carpenter (1976), Appendix E.

a high of .53 for the DBD ODD scale. Many correlations were statistically significant, but agreement was modest at best, with most correlations falling into the .30s and .40s range. The Pearson correlations, while slightly less variable in magnitude, were similarly low to moderate in size, and about half of them were statistically significant. With the exception of the DBD CD scale, Pearson correlations between dimensional scores were generally in the .40–.50 range, which is slightly higher than the intraclass correlations that consider disparities in scores, not just relative ranking. Agreement between two randomly selected teachers was nearly identical to the averaged Pearson correlations with the exception of agreement for CD, which was much higher, presumably because it was not weighted by correlations calculated from increasingly smaller numbers of teachers.

Agreement for Teachers' Ratings Coded to Reflect Deviant Versus Normal Functioning

Table 3 shows percentage agreement when teachers' ratings were categorically coded to reflect normal versus deviant functioning. These analyses were conducted for two, and then for three, randomly selected teachers for each student. Columns A through D show the agreement frequencies for two teachers: Column A shows the numbers of adolescents who were classified as behaviorally deviant because they reached the clinical range on the TRF or IOWA or because they met DSM–IV criteria on the DBD, as rated by both teachers; column B shows the numbers of adolescents who were classified as behaviorally deviant based on only one teacher's rating; column C shows the remaining adolescents who did not meet these criteria by either teacher's report; and column D shows, for each of the questionnaires, the total number of adolescents who were rated by two teachers.

The percentage of agreement between the two teachers' ratings for presence or absence of deviance, shown in column E of Table 3, shows that two teachers' ratings converged for 68% to 96% of adolescents. When agreement across three randomly selected teachers was examined, shown in column F of Table 3, convergence decreased on average by about eight percentage points. Agreement for three teachers ranged from 52% to 88% of adolescents. These figures suggest moderate to high levels of agreement when comparing teachers' ratings in terms of categorical diagnostic severity. However, in Table 4, the corresponding kappas and intraclass correlations are shown. They indicate that agreement above that expected by chance was not particularly strong; interrater reliability ranged from a low of .17 to a high of .48 for two teachers, and it ranged from a low of .30 to a high of .57 for three teachers. Furthermore, these indices at best only approached what are considered to be acceptable levels of reliability (e.g., kappa greater than or equal to .60, Hartmann, 1977).

Examination of the simple frequencies in columns A through D of Table 3 showed that these moderate agreement figures were driven by large numbers of adolescents who were rated by both teachers as having no clinically significant problems, Between 55% and 95% of adolescents were rated by two teachers as functioning below clinical threshold criteria, depending upon the particular subscale of interest. Consequently, we examined percentage agreement for the pool of adolescents for whom at least one teacher identified significant problems by virtue of high ratings. This procedure was used to compare results directly to those reported by Simpson (1991). The last column in Table 3 shows that, when examined for two randomly selected teachers, between 17 and 38 percent of adolescents had two

teachers who agreed that problems were significant. Thus, agreement was quite low when considering this relatively small but especially high-risk group.

Table 4 also allows comparison of agreement between teachers when DBD scores are used to assign DSM-IV diagnoses of ADHD Inattentive Type, ADHD Hyperactive-Impulsive Type, ADHD Combined Type, ODD, and CD. Agreement was slightly lower for ADHD Inattentive Type (approximately .30 vs. .40–.50 for the other disruptive behavior disorders). This difference reflects the frequencies in Table 3, which show a much larger proportion of adolescents (20 out of 66) with teacher disagreements for ADHD Inattentive Type (column B) than for the other ADHD subtypes or disruptive behavior disorders.

Pearson Correlations Among the Three Behavior Rating Scales

Table 5 shows correlations between the dimensional subscale scores for the TRF, IOWA, and DBD. These were calculated using a single randomly selected teacher for each adolescent. As can be readily seen, all correlations were medium to large in size and all were statistically significant at $p < .001$ or better. Correlations among the subscales assessing the ADHD symptoms of inattention, impulsivity, and overactivity (TRF Attention, IOWA IO, DBD Inattention, and DBD Impulsivity-Overactivity) ranged from a low of .55 (between IOWA IO and DBD Impulsivity-Overactivity) to a high of .83 (between DBD Attention and DBD Impulsivity-Overactivity). The two lowest correlations between these subscales, .56 and .55, were between the DBD Impulsivity-Overactivity and the TRF Attention and between the DBD Impulsivity-Overactivity and the IOWA IO subscales, respectively, which probably reflects the relatively lower frequency of impulsivity items on the TRF and IOWA subscales (1 item each) than on the DBD subscale (3 items). Correlations among the subscales assessing oppositional behavior (TRF Aggression, IOWA OD, and DBD ODD) ranged from .63 to .79. The correlation between the two subscales assessing delinquent behavior (TRF Delinquency, DBD CD) was quite high ($r = .85$).

Discussion

There are good reasons to expect that agreement among teachers' ratings should be lower for adolescents than for children. Developmental and contextual shifts occur between childhood and adolescence that could easily affect impressions, and expressions, of adolescent behavior. This study showed that when two or more teachers' ratings of adolescents are compared, there is indeed a high likelihood that they will be different in relative ranking and in absolute level. We found, for several well-known standardized measures, that Pearson correlations between different teachers' behavioral ratings were often statistically significant but only moderate in size. Most Pearson correlations were in the .40–.50 range. As expected, however, intraclass correlations were lower than the Pearson correlations and generally ranged from .20–.50. Thus, this study found that behavior ratings by secondary school teachers are likely to differ in two ways. Not only is a student's behavior likely to be ranked differently from one teacher to the next, but the absolute value of the score the student receives is likely to vary.

Agreement between teachers in identifying children as behaviorally deviant (i.e., scoring above research cutoff scores or in the diagnostic range) depended upon the index of

agreement under examination. Percentage agreement indices for two and for three teachers was in the good to excellent range, with figures ranging from 68% to 96% for two teachers and from 52% to 88% for three teachers. However, percentage agreement indices, while highly descriptive, are known to be inflated because of their failure to consider chance levels of agreement (Bartko & Carpenter, 1976; Shrout & Fleiss, 1979). When more conservative measures of agreement were used (the intraclass correlation coefficient and kappa), agreement dipped to a low of. 17 and to a high of .48 for two teachers and to a low pf .30 and to a high of .57 for three teachers. Thus, our findings indicated that agreement above that expected by chance is poor when teachers' ratings are used to discern whether a student's behavior is sufficiently impaired as to warrant a diagnosis or clinical attention.

Although many of our agreement indices met criteria for statistical significance, their magnitudes were low. Furthermore, although the addition of a third teacher slightly increased agreement (Table 4), the correlations remained low. In fact, most of the Pearson correlations were lower than the average Pearson correlation between teacher reports of .64 reported by Achenbach et al. (1987) in their well-known meta-analytic review of cross-informant correlations for children. The discrepancy between findings most likely rests with sample differences between our study and those in the Achenbach review. Of the 20 teacher-teacher correlations cited by Achenbach and colleagues, only three represented agreement for adolescents (r = .57 for 6- to 16-year-old outpatients; Achenbach & Edelbrock, 1986; r = .46 and r = .44 for regular classroom seventh and eighth graders, respectively; Quay & Quay, 1965). Quay and Quay was the only study to examine agreement for a purely adolescent sample. Indeed, Achenbach and colleagues assert that the size of the correlations obtained in their review were affected somewhat by the age of the participants and types of problems, with higher correlations obtained for ratings of 6- to 11-year olds and undercontrolled problems. Our findings bolster this conclusion, and they concur with those of Quay and Quay and more recently with Simpson (1991), to suggest that correlational agreement among teachers' behavior ratings is only modest at best. Our findings further add to those of previous studies of normal children by showing that agreement is also poor for students with a history of disruptive behavior problems and that agreement does not appear to be affected by the choice of measure (at least not by those used in this study).

Our findings have implications for diagnostic reliability and, consequently, assessment procedures for the disruptive behavior disorders. Current estimates of reliability for *DSM-IV* childhood disruptive behavior disorders range from kappas in the .50s (Lahey, Applegate, Barkley, et al., 1994; Lahey, Applegate, McBurnett et al., 1994) to kappas in the .80s (Morgan, Hynd, Riccio, & Hall, 1996). However, these figures pertain to agreement involving multiple-reporter structured interviews and diagnoses by clinicians who have access to full assessment batteries. Our findings suggest that when a single assessment source (in this case, secondary school teacher reports in the form of standardized rating scales) is used to generate a diagnosis of a disruptive behavior disorder, that agreement is much lower than that suggested by previous research. In fact, it is well below the threshold kappa of .60 suggested by Hartmann (1977). In order to reach acceptable levels of diagnostic consistency, clinicians should consider multiple sources of information (such as parent and teacher reports using standardized rating scales and interview information, previous treatment, and school records). This assessment recommendation appears especially

important for the *DSM-IV* inattentive subtype of ADHD, which had the lowest agreement, and which may increase in prevalence for adolescents. Many more adolescents in our study met criteria for the inattentive subtype than for the hyperactive-impulsive and combined subtypes based on teacher DBD ratings. Given the relatively subtle expression of inattention symptoms and the overall less extreme presentation of ADHD symptoms in adolescence (Evans, Vallano, & Pelham, 1995), clinicians should give consideration to a full range of assessment information when establishing a diagnosis.

Clearly, there is considerable variability among teachers in their perceptions of student behavior, even when ratings are of overt behavioral difficulties such as in this study (e.g., oppositionality as opposed to low self-esteem). However, we do not suggest that this variability weakens the utility of collecting such ratings. Rather, we strongly suggest that researchers and clinicians should respect this variability by incorporating multiple teacher reports into assessment batteries. Clinical assessment may be deficient when ratings are collected from only one secondary school teacher, particularly when behavioral problems are suspected. In fact, agreement was quite low when examined for adolescents for whom at least one teacher had identified significant problems; agreement percentages were in the 20s and 30s, and even these are likely to be inflated. These agreement statistics are nearly identical to those found by Simpson (1991) for a randomly selected sample of high school students. Therefore, agreement may not be affected by the psychiatric history of the sample. However, our findings pertain to a sample of adolescents with a childhood history of ADHD whose functioning at follow-up was variable. Whether agreement would be different for a sample of adolescents currently seeking treatment, or for adolescents with different types of mental health problems, requires further research.

We hypothesize that the substantial variability in ratings for all measures is a function of adolescent and teacher intrapersonal, interpersonal, and contextual domains that warrants assessment in future research. For example, little research has been conducted to systematically examine the impact of goodness-of-fit between student and teacher on teacher impressions of behavior (Greene, 1996). Multiple teacher variables, such as flexibility of teacher expectations (Lloyd, Kauffman, Landrum, & Roe, 1991), knowledge of ADHD (Greene, 1996), interpretation of behavior and tolerance for misbehavior (Whalen, 1989), and behavior management practices (Good & Brophy, 1991) may lead to variability in ratings of students. Furthermore, negative halo effects may cause spuriously high ADHD ratings in oppositional children (Abikoff, Courtney, Pelham, & Koplewicz, 1993), and unrealistic expectations regarding Ritalin efficacy for adolescents may bias treatment decisions (Smith, Pelham, Gnagy, & Bukstein, 1998). Finally, student aptitude and preference for one subject over another, as well as classroom size and time of day, may further impact expression of adolescent behavior. Only research designed to systematically separate these sources of teacher rating variability can determine which factors most strongly affect teachers' decisions to endorse an item on a paper-and-pencil measure.

Our findings should not be taken as an endorsement of one measure over another for general clinical assessment. In fact, in spite of the different purposes for which these measures were initially developed (TRF for broad mental health assessment; IOWA for brief assessment of hyperactivity and aggression; DBD for diagnostic assessment of ADHD, ODD, and CD), the

correlations among them were all statistically significant and generally moderate to strong in magnitude. Thus, with the exception of ADHD symptoms (particularly impulsivity), there is considerable overlap in content domain between these measures. Furthermore, there are other measures available for consideration that were not the focus of this study (e.g., the Conners ASQ, Goyette et al., 1978; the Revised Behavior Problem Checklist, Quay & Peterson, 1983, 1984). Rather, our findings suggest that regardless of the measure used, ratings from more than one teacher and data from multiple sources should be collected when adolescents are the population of interest, whether for clinical or for research purposes.

Acknowledgments

This research was supported by National Institute of Alcohol Abuse and Alcoholism Grants AA07453, AA0626, and K21AA-00202; NIDA Grant DA05605; and NIMH Grants MH47390, MH4815, MH45576, MH50467, and MH533554.

References

Abikoff H, Courtney M, Pelham WE Jr, Koplewicz HS. Teachers' ratings of disruptive behaviors: The influence of halo effects. Journal of Abnormal Child Psychology. 1993; 21:519–533. [PubMed: 8294651]

Achenbach, TM. Integrative Guide for the 1991 CBCL14–18, YSR, and TRF profiles. Burlington: University of Vermont, Department of Psychiatry; 1991.

Achenbach, TM.; Edelbrock, C. Manual for the Teacher's Report Form and Teacher Version of the Child Behavior Profile. Burlington: University of Vermont, Department of Psychiatry; 1986.

Achenbach TM, McConaughy SH, Howell CT. Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. Psychological Bulletin. 1987; 101:213–232. [PubMed: 3562706]

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 3rd ed.. Washington, DC: Author; 1980.

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 3rd ed.. Washington, DC: Author; 1987. rev.

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed.. Washington, DC: Author; 1994.

Atkins MS, Pelham WE, Licht MH. The differential validity of teacher ratings of inattention/overactivity and aggression. Journal of Abnormal Child Psychology. 1989; 17:423–435. [PubMed: 2794255]

August GJ, Garfinkel BD. Comorbidity of ADHD and reading disability among clinic-referred children. Journal of Abnormal Child Psychology. 1990; 18:29–46. [PubMed: 2324400]

Barkley, RA. Attention deficit hyperactivity disorder. A handbook for diagnosis and treatment. New York: Guilford; 1990.

Barkley RA, Fischer M, Edelbrock CS, Smallish L. The adolescent outcome of hyperactive children diagnosed by research criteria: 1. An 8-year prospective follow-up study. Journal of the American Academy of Child and Adolescent Psychiatry. 1990; 29:546–557. [PubMed: 2387789]

Bartko JJ, Carpenter WT. On the methods and theory of reliability. The Journal of Nervous and Mental Disease. 1976; 163:307–317. [PubMed: 978187]

Evans, S.; Vallano, G.; Pelham, WE. Attention-deficit hyperactivity disorder. In: Van Hasselt, VB.; Hersen, M., editors. Handbook of adolescent psychopathology. A guide to diagnosis and treatment. New York: Lexington; 1995. p. 589-617.

Good, TL.; Brophy, JE. Looking in classrooms. 5th ed.. New York: HarperCollins; 1991.

Goyette CH, Conners CK, Ulrich RF. Normative data on revised Conners Parent and Teacher Rating Scales. Journal of Abnormal Child Psychology. 1978; 6:221–236. [PubMed: 670589]

Greenbaum PE, Dedrick RF, Prange ME, Friedman RM. Parent, teacher, and child ratings of problem behaviors of youngsters with serious emotional disturbances. Psychological Assessment. 1994; 6:141–148.

Greene, RW. Students with attention-deficit hyperactivity disorder and their teachers: Implications of a goodness-of-fit perspective. In: Ollendick, TH.; Prinz, RJ., editors. Advances in clinical child psychology. Vol. 18. New York: Plenum; 1996. p. 205-230.

Hartmann DP. Considerations in the choice of interobserver reliability estimates. Journal of Applied Behavior Analysis. 1977; 10:103–116. [PubMed: 16795538]

Hutton JB, Dubes R, Muir S. Assessment practices of school psychologists: Ten years later. School Psychology Review. 1992; 21:271–284.

Lahey BB, Applegate B, Barkley RA, Garfinkel B, McBurnett K, Kerdyk L, Greenhill L, Hynd GW, Frick PJ, Newcorn J, Biederman J, Ollendick T, Hart EL, Perez D, Waldman I, Shaffer D. DSM-IV field trials for oppositional defiant disorder and conduct disorder in children and adolescents. American Journal of Psychiatry. 1994; 151:1163–1171. [PubMed: 8037251]

Lahey BB, Applegate B, McBurnett K, Biederman J, Greenhill L, Hynd GW, Barkley RA, Newcorn J, Jensen P, Richters J, Garfinkel B, Kerdyk L, Frick PJ, Ollendick T, Perez D, Hart EL, Waldman I, Shaffer D. DSM–IV field trials for attention deficit hyperactivity disorder in children and adolescents. American Journal of Psychiatry. 1994; 151:1673–1685. [PubMed: 7943460]

Lee SW, Elliott J, Barbour JD. A comparison of cross-informant behavior ratings in school-based diagnosis. Behavioral Disorders. 1994; 19:87–97.

Lloyd JW, Kauffman JM, Landrum TJ, Roe DL. Why do teachers refer pupils for special education? An analysis of referral records. Exceptionality. 1991; 2:115–126.

Loney, J.; Milich, R. Hyperactivity, inattention, and aggression in clinical practice. In: Wolraich, M.; Routh, DK., editors. Advances in developmental and behavioral pediatrics. Vol. 3. Greenwich, CT: JAI; 1982. p. 113-147.

McConaughy SH. Advances in empirically based assessment of children's behavioral and emotional problems. School Psychology Review. 1993; 22:285–297.

Morgan AE, Hynd GW, Riccio CA, Hall J. Validity of *DSM–IV* ADHD predominantly inattentive and combined types: Relationship to previous DSM diagnoses/subtype differences. Journal of the American Academy of Child and Adolescent Psychiatry. 1996; 35:325–333. [PubMed: 8714321]

Pelham WE, Evans S, Gnagy E, Greenslade KE. Teacher ratings of *DSM–III–R* symptoms for the disruptive disorders: Prevalence, factor analyses, and conditional probabilities in a special education sample. School Psychology Review. 1992; 21:285–299.

Pelham WE, Gnagy EM, Greenslade KE, Milich R. Teacher ratings of *DSM–III–R* symptoms for the disruptive behavior disorders. Journal of the American Academy of Child and Adolescent Psychiatry. 1992; 31:210–218. [PubMed: 1564021]

Pelham WE, Milich R, Murphy DA, Murphy HA. Normative data on the IOWA Conners Teacher Rating Scale. Journal of Clinical Child Psychology. 1989; 18:259–262.

Phares V, Compas BE, Howell DC. Perspectives on child behavior problems: Comparisons of children's self-reports with parent and teacher reports. Psychological Assessment. 1989; 1:68–71.

Pillow DR, Pelham WE, Hoza B, Molina BSG, Stultz CH. Confirmatory factor analyses examining attention deficit hyperactivity disorder symptoms and other childhood disruptive behaviors. Journal of Abnormal Child Psychology. 1998; 26:293–309. [PubMed: 9700521]

Quay, HC.; Peterson, DR. Interim manual for the Revised Behavior Problem Checklist. University of Miami; 1983. Unpublished manuscript

Quay, HC.; Peterson, DR. Appendix 1 to the interim manual for the Revised Behavior Problem Checklist. University of Miami; 1984. Unpublished manuscript

Quay HC, Quay LC. Behavior problems in early adolescence. Child Development. 1965; 36:215–220. [PubMed: 14296788]

Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin. 1979; 86:420–428. [PubMed: 18839484]

Simpson RG. Agreement among teachers of secondary students in using the Revised Behavior Problem Checklist to identify deviant behavior. Behavioral Disorders. 1991; 17:66–71.

Smith, B.; Pelham, W.; Gnagy, E.; Bukstein, O. Long-acting versus short-acting stimulant treatment of adolescents with ADHD: A preliminary study in natural settings. 1998. Manuscript in preparation.

Spitzer RL, Cohen J, Fleiss JL, Endicott J. Quantification of agreement in psychiatric diagnosis. Archives of General Psychiatry. 1967; 17:83–87. [PubMed: 4952165]

Sprague, RL.; Christensen, DE.; Werry, JS. Experimental psychology and stimulant drugs. In: Conners, CK., editor. Clinical use of stimulant drugs in children. Amsterdam: Excerpta Medica; 1974. p. 141-164.

Whalen, CK. Attention deficit and hyperactivity disorders. In: Ollendick, TH.; Hereon, M., editors. Handbook of child psychopathology. 2nd ed.. New York: Plenum; 1989. p. 131-169.

Table 1

Number of Adolescents for Whom Teachers Completed Each of the Three Questionnaires

	TRF	IOWA	DBD
No. of Teachers			
5	14	18	18
4	19	19	21
3	17	17	17
2	13	11	11
1	9	8	11
Total No. of Adolescents With One or More Questionnaires	72	73	78
Total No. of Questionnaires	232	247	258

Note: TRF = Teacher Report Form; IOWA/Conners = IOWA/Abbreviated Conners; DBD = Disruptive Behavior Disorders Rating Scale.

Table 2

Agreement Between Teachers for Dimensional Subscale Scores

Questionnaire	Intraclass Correlation Across All Teachers	Average Pearson Correlation Across All Teachers	Pearson Correlations for Only Two Teachers
TRF			
Attention	.51*	.46*	.48*
Aggression	.32	.44	.51*
Delinquency	.41*	.53*	.53*
IOWA			
IO	.35*	.35	.37*
ODD	.21	.49*	.47*
DBD			
Inattention	.48*	.40	.42*
Impulsivity-Overactivity	.46*	.39	.43*
ODD	.53*	.49*	.50*
CD	.13	.23	.49*

Note: TRF = Teacher Report Form; IOWA = IOWA/Abbreviated Conners; IO = inattention/overactivity; ODD = oppositional defiant disorder; DBD = Disruptive Behavior Disorders Rating Scale; CD = conduct disorder. Pearson correlations are transformed back from the Fisher *z* scale. With the exception of DBD CD ($df = 15$), *df* for the average Pearson correlations ranged from 29 to 35 (an average of the *df* across correlations used to calculate an average correlation).

* $p < .01$ or better.

Table 3

Percentage Agreement Between Teachers for Normal Versus Deviant Functioning

Questionnaires	A	B	C	D	E[a]	F[b]	G[c]
TRF							
Attention	4	16	43	63	75	68	20
Aggression	4	10	49	63	84	80	29
Delinquency	2	10	51	63	84	88	17
IOWA							
IO	7	20	36	63	68	56	26
ODD	8	13	43	64	80	74	38
DBD							
ADHD Inattentive Type	10	20	36	66	70	52	33
ADHD Hyperactive-Impulse Type	2	7	57	66	89	84	22
ADHD Combined Type	2	5	59	66	92	86	29
ODD	4	8	54	66	88	82	33
CD	1	2	53	56	96	86	33

Note: TRF = Teacher Report Form; IOWA = IOWA/Abbreviated Conners; IO = inattention/overactivity; ODD = oppositional defiant disorder; DBD = Disruptive Behavior Disorders Rating Scale; ADHD = attention deficit hyperactivity disorder. Percentage agreement represents similarity across teachers for scoring above cutoff on a dimensional score (TRF or IOWA) and for meeting *DSM-IV* diagnostic criteria on the DBD. With the exception of column F, two teachers were randomly selected for each adolescent. Column A indicates those adolescents who were rated as deviant by both teachers. Column B indicates those who were rated as deviant by only one of the teachers. Column C indicates those who were not rated as deviant by either teacher. Column D = A + B + C.

[a] Agreement for presence or absence of problems as rated by two teachers; [(A + C) / D] × 100.

[b] Agreement for presence or absence of problems as rated by 3 teachers; frequencies used to calculate these percentages are not provided in table.

[c] Percentage agreement for presence of problems as rated by two teachers; [A / (A + B)] × 100.

Table 4

Extent of Agreement Between Teachers for Normal Versus Deviant Functioning: Kappa and Intraclass Correlations

Questionnaires	Kappa for Presence or Absence of Problems as Rated by 2 Teachers	Intraclass Correlation for Presence or Absence of Problems as Rated by 3 Teachers
TRF		
Attention	.18	.33*
Aggression	.36*	.45*
Delinquency	.21	.41*
IOWA		
IO	.17	.31*
ODD	.43*	.57*
DBD		
ADHD Inattentive Type	.28	.30*
ADHD Hyperactive-Impulsive Type	.31	.38*
ADHD Combined Type	.40*	.41*
ODD	.43*	.52*
CD	.48*	.52*

Note: TRF = Teacher Report Form; IOWA= JOWA/Abbreviated Conners; IO = inattention/overactivity; ODD = oppositional defiant disorder; DBD = Disruptive Behavior Disorders Rating Scale; ADHD = attention deficit hyperactivity disorder; CD = conduct disorder. Agreement represents similarity across teachers for scoring above cutoff on a dimensional score (TRF or IOWA) and for meeting *DSM-IV* diagnostic criteria on the DBD. To correct for Type 1 error, statistical significance is only indicated at $p < .01$ or better.

*$p < .01$ or better.

Table 5

Correlations Among the Three Behavior Rating Scales

	TRF			IOWA/Conners		DBD			
Questionnaire	**Attention**	**Aggression**	**Delinquency**	**IO**	**ODD**	**Attention**	**Impulsivity**	**ODD**	**CD**
TRF									
Attention	10.55								
Aggression	.66	11.00							
Delinquency	.61	.81	3.06						
IOWA									
IO	.61	.57	.54	3.86					
ODD	.49	.79	.70	.65	4.34				
DBD									
Inattention	.74	.56	.52	.67	.41	.79			
Impulsivity	.56	.72	.54	.55	.46	.83	.75		
ODD	.54	.78	.65	.51	.63	.71	.87	.69	
CD	.47	.72	.85	.46	.47	.62	.63	.77	.63

Note: TRF = Teacher Report Form; IOWA/Conners = IOWA/Abbreviated Conners Teacher Rating Scale; DBD = Disruptive Behavior Disorders Scale; IO = inattention/overactivity; ODD = oppositional defiant disorder; CD = conduct disorder. Correlations were calculated using one randomly selected teacher for each student. Standard deviations are on the diagonal. Sample sizes ranged from 73 to 78. All correlations are statistically significant at $p < .001$ or better.