

Measuring Symptoms and Functioning of Youth with ADHD in Middle Schools

Steven W. Evans,¹ Jessica Allen,¹ Sheryle Moore,¹ and Victoria Strauss¹

Received July 24, 2003; revision received March 21, 2004; accepted June 21, 2004

The identification of reliable and valid means for evaluating the effectiveness of school-based treatments and completing diagnostic evaluations of middle school aged students are needed. The present study examined the inter-rater agreement of teacher ratings and the relationship between ratings and observational data in a middle school setting. The data are interpreted in the context of differences between a secondary and elementary school setting. Teacher ratings and observational data were collected regularly over the course of two academic years for middle school students diagnosed with ADHD. The results indicate low rates of inter-rater agreement as well as low rates of agreement between teachers and observational data, and between observational data collected in different classrooms. Inter-rater agreement was lowest in late fall and gradually increased over the second half of the year. Implications for conducting treatment outcome evaluations of school-based treatment programs and diagnostic evaluations are discussed.

KEY WORDS: schools; measurement; reliability; adolescence; ADHD.

The development of effective school-based treatments for children and adolescents relies on the availability of reliable and valid measures. These measures are needed to diagnose children presenting for services and to evaluate outcomes to determine the effects of the treatment. Recent guidelines for the evaluation and treatment of youth with attention-deficit/hyperactivity disorder (ADHD) by the American Academy of Pediatrics (2000) and the American Academy of Child and Adolescent Psychiatry (1997) recommend collecting ratings from teachers to guide diagnostic decisions. Teacher ratings are valued since teachers see the children in a structured setting and have a sample of normal-functioning children to which they can compare the child being assessed. The American Academy of Pediatrics noted that teacher ADHD-specific questionnaires have demonstrated strong sensitivity and specificity such that they can accurately differentiate between children with and without the disorder.

In addition to their diagnostic utility, teacher ratings are also widely used to measure the outcomes associated with treatment. This is especially true for school-based treatment (e.g., Evans, Axelrod, & Langberg, 2004; Reid, Eddy, Fetrow, & Stoolmiller, 1999). Studies frequently use teacher ratings in a repeated measures design in order to draw conclusions about the effect of the interventions being tested.

Although the literature on and use of teacher ratings for elementary school students is extensive, there are no empirically supported recommendations for their use in secondary school contexts. One of the first dilemmas faced by a clinician attempting to collect rating scale data from a middle or high school is selecting a teacher to ask to complete the ratings. Achenbach (1991, p. 109) has recommended giving the rating scale to “whichever teachers know the child reasonably well.” This criterion is difficult to operationalize in a secondary school setting where a student might have six teachers for approximately one hour per day and some teachers might only have the student in a class for one semester. In the research literature on adolescents, this problem is frequently not directly addressed. Research studies that employ teacher ratings of adolescents tend to use only one teacher’s ratings and

¹Department of Psychology, James Madison University, Harrisonburg, Virginia.

²Address all correspondence to Steven W. Evans, Ph.D., MSC 9013, James Madison University, Harrisonburg, VA 22807; e-mail: evanssw@jmu.edu.

provide little information on the rationale for selecting the teacher (Greenbaum, Dedrick, Prange, & Friedman, 1994; Lee, Elliott, & Barbour, 1994; Phares, Compas, & Howell, 1989), whereas other investigators do not report having collected any data directly from schools as part of their evaluation procedures for ADHD with adolescents (Barkley, Edwards, Laneri, Fletcher, & Metevia, 2001; Biederman et al., 1999). Furthermore, the investigators conducting the DSM-IV field trials for ADHD collected teacher data for the elementary school students in their study, but did not collect teacher data for the subjects in secondary schools due to "the relative lack of close contact between students and teachers in middle and high schools." (p. 1674; Lahey et al., 1994). Finally, Pliszka et al. (2003) reported that teacher rating data were difficult to collect and did not appear to be valued by the psychiatrists conducting medication trials.

Avoiding the collection of teacher rating data to identify the level of school functioning and symptoms would be appropriate if there were other viable sources for this information. Some have argued that teacher rating data are critical. Jensen et al. (1999) noted in a study of the evaluation of psychopathology in youth ages 9 to 17 that "information from a reliable school informant is a sine qua non, even in the presence of parent ADHD symptom endorsement (or absence)." (p. 1575). There are recent data that support this assertion. Although it is possible that collecting evaluation data only from parents is a viable method for accurately evaluating school functioning and symptoms, there is evidence that parents and teachers provide different information. Mitsis, McKay, Schulz, Newcorn, and Halperin (2000) reported that parent-teacher agreement on the diagnosis of ADHD was moderate (74%). Teachers tended to report a greater number of school symptoms than parents and parent reports of school behavior were found to be influenced by their observation of behavior in the home. The authors concluded that "parent reports of ADHD behaviors in school are not an adequate substitute for direct teacher input." (p. 312). Given that participants in this study were elementary school students and parents often profess to know more about their children's school behavior at the elementary level than at the middle or high school level, parent reports of school behavior in secondary schools might be less adequate than they were found to be with younger children.

Because school functioning is a high priority and frequently the reason for referral among adolescents with ADHD, it is central to diagnosis and to the evaluation of treatment. Hence, it is critical to understand the strengths and weaknesses of measures of school symptoms and functioning. Several studies have been completed to

determine the degree of agreement between teachers on rating scale information, many of which were summarized by Achenbach, McConaughy, and Howell (1987). In their meta-analysis, data were pooled from 119 studies investigating inter-rater reliability on variables ranging from subtle to readily observable behaviors. They found that the mean correlation between teachers' ratings was .64, suggesting that although there is moderate consistency between the ratings, there is also significant disagreement. They also found that the mean correlation between teachers' ratings and observers' ratings was .42, again indicating a significant amount of discrepancy between teachers and trained observers' accounts of student behavior. It is important to note that the sample consisted primarily of elementary school aged students and their teachers. Achenbach et al. (1987) noted that higher correlations were found when comparing teachers' ratings of 6- to 11-year olds (elementary school students), or when those ratings assessed externalizing behavior problems, than when comparing teachers' ratings of secondary school students, or of internalizing behavior problems. These findings suggest that secondary school teachers provide less reliable ratings than do their elementary school counterparts, possibly due to differences between settings, student/teacher time together, and student/teacher ratios.

In a study conducted in a secondary school setting, Simpson (1991) investigated teachers' agreement on ratings of deviant secondary school students. Results showed low to moderate relationships among teacher ratings with only one correlation coefficient indicating substantial agreement (.54 for conduct disorder/girls). In another study, Molina, Pelham, Blumenthal, and Galiszewski (1998) investigated the agreement between secondary school teachers' ratings of adolescents with a childhood history of ADHD using intraclass correlations. Their findings indicated minimal to substantial agreement between multiple teacher ratings, ranging from .13 to .53. The authors concluded that ratings by secondary school teachers are likely to vary both in relative and absolute value. They suggested that these findings support the need to collect rating scale data from multiple teachers. Although this conclusion seems warranted, their data raise questions about how to interpret diverse rating scores from multiple teachers.

A recent report by Evans, Langberg, Raggi, Allen, and Buvinger (2004) indicated that the use of data from multiple teachers makes treatment outcome difficult to determine. Whereas parent report data revealed fairly consistent benefits from the school-based treatment program, teacher data were very inconsistent and difficult to interpret. Evans et al. (2004) reported some effect sizes in opposite directions with a magnitude greater than one for

the same child as rated by different teachers, indicating that more data do not always improve clarity. Methods for integrating data from multiple informants have been proposed to address this inconsistency (Kraemer et al., 2003), including collecting data from different perspectives and contexts and using principal-component analyses to identify a trait gold-standard. Kraemer et al. (2003) also emphasized that the selection of informants is critical and suggested that multiple informants from the same context or perspective would provide little benefit to the identification of true characteristics since their data would be highly correlated. The findings reported in the previously reviewed studies suggest that data from multiple secondary school teachers are not highly correlated and that secondary school teachers might represent a variety of unique contexts (classrooms) and perspectives (different teachers).

In addition to being subject to many reporter biases known to influence teacher ratings (Abikoff, Courtney, Pelham, & Koplewicz, 1993; Chang & Sue, 2003; Merrel, 2000), secondary school teachers also face other limitations because they work in a very different setting from elementary school teachers. Elementary school teachers observe their students in a wide range of academic and social situations including classwork, walks to the restrooms, and play activities. Furthermore, since elementary school teachers are seen as the focal point for their assigned children, other teachers, staff, and administrators keep them updated about events outside of the classroom such as behavior on the school bus, progress in physical education class, and behavior in the lunchroom. In addition, parents tend to communicate more with elementary school teachers than with secondary school teachers. As a result, elementary school teachers develop considerable breadth and depth of knowledge about the 20–30 students with whom they spend approximately 6 hr per day. In contrast, secondary school teachers are frequently assigned over 100 children and observe their behavior for approximately 1 hr per day. Secondary teachers are not the focal point for their assigned students' education, so it is common that they do not know about problems in settings outside their classrooms. Finally, their communication with most children's parents often is minimal to nonexistent. As a result, the reliability and validity of secondary school teachers as sources of information about students is likely to be diminished compared to that of elementary school teachers.

These problems have led many to abandon the practice of obtaining teacher rating data from secondary schools, but no viable alternative currently exists and the assessment of school functioning and symptom manifestation is critical to the diagnostic process and evaluation of

treatment benefits. New assessment procedures are needed and should take into account the unique characteristics of assessment in secondary schools. Specific questions need to be addressed in order to guide the development of such procedures.

To begin, additional information is needed that describes the behaviors and areas of functioning that secondary school teachers are best able to report. Because elementary school teachers observe a broad range of student activities they might be able to provide reliable and valid data about specific areas of functioning such as peer relations and self-esteem, whereas secondary school teachers might not. In addition, our understanding of the assessment process in secondary schools would be improved if we understood how teachers' cumulative acquisition of knowledge about students influences their ability to rate behavior. Low rates of agreement reported by other investigators might reflect very low rates of agreement early in the year along with good rates of agreement later. Finally, it is important to understand if different teachers do, in fact, represent different contexts, because this has an impact on the selection of raters (Kraemer et al., 2003). Findings reported by Simpson (1991) and Molina et al. (1998) indicate that different teachers might represent different contexts in secondary schools. If this is true, then there are important implications for methods of collecting teacher rating data as well as for diagnostic decision making.

METHOD

Participants

The sample consisted of secondary-school students from a regular education public middle school in Rockingham County, Virginia. The students were involved in the Challenging Horizons Program (CHP), a comprehensive school-based psychosocial treatment program for adolescents with ADHD. A prerequisite for enrollment in the CHP was a diagnosis of ADHD (Combined, Inattentive, or Hyperactive/Impulsive Type) based on a comprehensive evaluation. After an initial outside referral to the program, parents and two of each student's teachers were asked to complete the Behavioral Assessment System for Children (BASC, Reynolds & Kamphaus, 1992), the Impairment Rating Scale (IRS, Fabiano & Pelham, 2002), and the ADHD Rating Scale-IV (DuPaul, Power, Anastopoulos, & Reid, 1998). In addition to these rating scales, one parent completed the sections of the Diagnostic Interview Schedule for Children (DISC-IV: Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000) pertaining to

ADHD, oppositional defiant disorder, conduct disorder, enuresis/encopresis, adjustment disorder, depression, anxiety, and bipolar disorder.

The sample consisted of 21 male students (84%) and 4 female students (16%). All of the participants were Caucasian and were between 11 and 14 years old. Eighteen students (72%) met diagnostic criteria for ADHD-Combined Type and seven students (28%) met diagnostic criteria for ADHD-Inattentive Type. Eleven of the students met diagnostic criteria for other disorders including recurrent major depressive disorder ($N = 1, 4\%$), conduct disorder ($N = 2, 8\%$), oppositional defiant disorder ($N = 6, 24\%$), enuresis ($N = 1, 4\%$), adjustment disorder ($N = 1, 4\%$), and mathematics disorder ($N = 1, 4\%$). After acceptance into the program, students completed the Wechsler Intelligence Scale for Children-Third Edition (WISC-III, Wechsler, 1991). Full scale IQ scores ranged from 77 to 118 ($M = 97.1, SD = 10.9$). All participants and their parents signed assent and consent forms approved by the Institutional Review Board prior to their participation in the research.

Each one of fifteen teachers involved in the study rated between 1 and 8 children over the course of two academic years. The mean number of students rated by each teacher was 3.93. Seven of the teachers rated 5 or more students with the remaining eight rating between 1 and 4 students. Six science teachers and nine math teachers were asked to complete ratings. In one case, only one teacher rating was obtained because that student was involved in an alternative-learning environment where he had only one teacher so these data were only used in the validity portion of the analyses. Science and math teachers were selected to obtain data from a variety of classroom settings. Math classes tend to include a considerable amount of seatwork whereas science classes frequently require students to be out of their seat and working in groups.

Measures

The ADHD Rating Scale and the IRS were administered monthly to teachers of students in the CHP. The observational measures were completed on a weekly basis.

ADHD Rating Scale-IV

The ADHD Rating Scale-IV (DuPaul et al., 1998) consists of 18 items derived from the diagnostic criteria set forth in the DSM-IV (American Psychiatric Association, 1994). Two factors have been identified: a 9-item Inattention factor and a 9-item Hyperactivity-Impulsivity factor, both of which follow from the two-dimensional

structure of ADHD specified by the DSM-IV. Each item is rated on a 4-point Likert scale with options including 0 (*never or rarely*), 1 (*sometimes*), 2 (*often*), and 3 (*very often*). Three scores are obtained from this measure: an inattention subscore, a hyperactivity subscore, and a total score (sum of the two subscores). The maximum score possible for either subscale is 27, yielding a maximum score of 54 for the total. Reliability and validity data are summarized in the manual for the assessment (DuPaul et al., 1998) and indicate that the teacher version has high internal consistency (alphas range from .88 to .96), good test-retest reliability (Pearson r 's range from .88 to .90), and that scores correlated significantly with other commonly used measures of ADHD symptoms (sharing between 53 and 77% of the variance with some measures). These reliability and validity data were obtained using a sample that included middle school age children.

Impairment Rating Scale

The Impairment Rating Scale (Fabiano & Pelham, 2002) was developed to assess the areas of functioning that are frequently problematic for children with ADHD. The school version of this measure consists of six items with responses ranging from *No Problem/definitely does not need treatment* to *Extreme Problem/Definitely needs treatment*. The areas of functioning assessed include peer relations, relationship with teachers, academic progress, classroom behavior, self-esteem, and overall functioning. Teachers are asked to mark an "X" on a line between both responses indicating the severity of the student's problems in each specific area. A scoring template is used to apply a quantitative score to the location of the teachers' marks and yields scores ranging from 0 to 6. Normative data on the teacher version have not been obtained for an adolescent population. The measure was selected because it efficiently assesses functioning across pertinent domains in adolescents with ADHD. Reliability and validity data are summarized in two reports by the authors (Fabiano & Pelham, 2002; Fabiano et al., 2004) and indicate that the teacher version has high internal consistency ($\alpha = .95$), good parent and teacher version 3 to 4 month test-retest reliabilities (Pearson r 's range from .74 to .96), and good convergent and discriminant validity (PPP = .90; NPP = .74).

Both the ADHD Rating Scale-IV and the Impairment Rating Scale were distributed to the students' mathematics and science teachers at the end of each month. The teachers were asked to complete the scales based on behavior exhibited during the previous month and return the form within one week. Participating teachers were compensated

for their time based on the number of students they rated in their classes.

Classroom Observations

In conjunction with teacher ratings, observational data were obtained from the regular education classrooms of the teachers who completed the rating scales. Classroom observation data were gathered using the CAT Classroom Observations Procedure (Pelham et al., 2001). The procedure was designed so that two 44-min observations per child per week were scheduled: one observation from the student's mathematics class and a second observation from their science class. Using this method, observers recorded on-task behavior, off-task behavior, and disruptive behaviors including aggression, verbal abuse, intentional destruction of property/inappropriate use of materials, cheating, interruption, talking to self, or being out of seat. The 44-min observation period was divided between one experimental participant and a control student. For the first minute (six 10-s intervals) the observer recorded whether the CHP student was on-task or off-task or engaged in any of the disruptive behaviors at the end of each 10-s interval. After 1 min, the observer began observing a control participant (i.e., a same-sex classmate who was not involved in the CHP). Although the CHP participant remained the same throughout the 44-min observation, the comparison child changed each minute.

RESULTS

Missing Data

One hundred ninety six of the possible 208 teacher ratings were obtained from the Science teachers (94.2%) and 186 of the possible 208 teacher ratings were obtained from the Math teachers (89.4%). Missing rating scales were due solely to teachers' failing to complete the rating scales during a given month. December was the month with the greatest amount of missing data, probably due to work load challenges associated with the winter break. Students were observed, on average, two to three times per month in each of their classes. Missing observational data were due largely to scheduling conflicts, students missing school due to illness and other extenuating circumstances, school-closures due to inclement weather, and altered class schedules as a result of additional school activities (i.e., assemblies, club meetings).

Inter-Observer Reliability

Observers were undergraduate students trained in the observation procedures. They were required to memorize the definitions of the observation categories and procedures, and then they went with a graduate supervisor to classrooms in a middle school to practice. Research assistants were trained to administer the observation system until they achieved phi-coefficients of .80 or greater in the training settings. The training data were not used in the analyses. During several of the actual observations, the graduate student in charge of observational data collection and the trained observers jointly observed the students. Phi-coefficients were calculated for each of these sessions and yielded a mean coefficient of .84. This score suggests adequate reliability and is at or above the level of the reliability coefficients reported in other research (e.g., Abikoff et al., 2002).

Agreement Between Teachers

As an index of agreement between teachers over the entire academic year, one-way random effect intraclass correlations (ICCs) for consistency were calculated for the Total, Inattention, and Hyperactivity subscores on the ADHD Rating Scale-IV, as well as for each of the six items on the Impairment Rating Scale. An overall measure of agreement was calculated (all months of data combined into one analysis). In addition, reliability rates were calculated for each individual month within the academic year. ICCs were chosen over Pearson correlations because they provide an index of the actual agreement between ratings from different teachers, as opposed to the index of association provided by Pearson correlations (Bartko & Carpenter, 1976). The single measure ICCs are reported because they give an index of agreement for one typical judge. The correlations found in these analyses are summarized in Table I. ICCs ranged from .06 on item 2 of the IRS (his or her relationship with the teacher) to .37 on the Hyperactivity subscale of the ADHD Rating Scale-IV. Analyses revealed significant agreement between teachers on each of the scores on the ADHD Rating Scale-IV and also on all items on the Impairment Rating Scale except item 2, $r = .06$, $p = .23$. Low teacher agreement also was found on the "self-esteem" and "relationship with peers" items, with higher rates of agreement found for the "academic progress" and "overall severity" items, as well as for scores on the ADHD Rating Scale-IV. It is important to interpret these rates of agreement in the context of indices of behavioral consistency between classrooms. Monthly percentages of on-task behavior were calculated across

Table I. Overall Teacher Agreement on ADHD Rating Scale-IV Scores and Individual IRS Items

Teacher rating	ICC	<i>n</i>
ADHD-RS total	.36***	176
ADHD-RS inattention	.30***	176
ADHD-RS hyperactivity	.37***	176
IRS relationship with peers	.23*	176
IRS relationship with teacher	.06	176
IRS academic progress	.37***	176
IRS effect on class	.28*	176
IRS self-esteem	.19*	167
IRS overall severity	.33***	174

Note. ADHD-RS: Attention Deficit Hyperactive Disorder-Rating Scale; IRS: Impairment Rating Scale; ICC: Intraclass Correlations.

* $p < .01$. ** $p < .001$. *** $p < .0001$.

observations within classroom and student and correlated with similar data from the other classroom. Agreement between these data were modest ($r = 0.25$; $p = .078$) suggesting that part of the variation in teacher ratings might be due to differences in rates of on-task behavior between classrooms.

These relatively low rates of agreement might have been influenced significantly by very low rates of agreement at the beginning of the year followed by gradual improvement as the teachers came to know the students. To test this theory the ICCs were calculated for each month to map the agreement in teachers' ratings over the course of the academic year (see Table II). Teachers agreed with one another on the ADHD Rating Scale-IV scores, and items 2, 3, and 6 of the IRS ("relationship with the teacher," "academic progress," and "overall severity and need for treatment") during the first month of the academic year, but in many cases began to show decreased rates of agreement as the first semester progressed (e.g., inattention subscore). The highest levels of agreement throughout the first semester were obtained for the hyperactivity subscore of the ADHD Rating Scale-IV. Teachers reached significant levels of agreement on the hyperactivity factor similar to those reported both by Molina et al. (1998) and Simpson (1991) in September, October, and November (ranging from .42 to .59). In December, teachers only reached significant levels of agreement with one-another on items 2, 3, and 4 of the IRS ("relationship with the teacher," "academic progress," and "how the student's problems affect the classroom in general"). During the months of January through April, teachers began to reach levels of agreement higher than those found for the first semester (except for September). Figure 1 provides a graphic representation of the monthly ICCs for the ADHD Rating Scale-IV. The items with the highest coefficients on the IRS (academic

progress and overall impairment) produced a pattern similar to that of the inattention factor of the ADHD Rating Scale.

Relationship to Observational Data

To determine whether or not teachers were rating the students in relation to behaviors manifested in the classroom (on-task behavior and disruptive behavior), Pearson correlations were calculated between the teachers' ratings and observational data obtained by trained classroom observers (see Table III). The correlations were calculated based on the ratings completed at the end of the month and the observational data collected during that same month. The observational data used in the analyses included observed on-task percentages and disruptive behavior percentages. The percentages indicate the portion of the 10-s observation intervals during which the student was either on-task or exhibiting disruptive behavior. Students' on-task behavior percentages and disruptive behavior percentages were correlated significantly with many of the teacher rating measures. On-task percentages were correlated significantly with all scores on the ADHD Rating Scale-IV and all items on the IRS. Correlations ranged from $-.195$ on the Inattention subscore of the ADHD Rating Scale to $-.34$ on item 4 of the IRS (how this child's problems affect your classroom in general). All correlations were negative and indicate that as students were more on-task in class, teachers rated these students as less impaired. Observed rates of disruptive behavior for the students were correlated significantly with all teacher ratings except the Hyperactivity subscore of the ADHD Rating Scale, item 2 of the IRS (his or her relationship with the teacher), and item 5 of the IRS (his or her self-esteem). Correlations ranged from .084 on item 2 (relationship with teacher) of the IRS to .245 on item 1 (his or her relationship with other children) of the IRS. The on-task measure correlated more highly with teacher ratings than did the disruptive behavior percentage, and there was very little variability in the magnitude of the correlations between the observational measures and specific teacher ratings suggesting little specificity in associations.

DISCUSSION

Three main questions were addressed in this study. The first sought to identify the areas of school functioning about which teachers are best able to report. The second question pertained to whether poor teacher agreement is due to teachers not knowing students well during the first few months of school, with agreement improving as they come to know their students better. The final question

Table II. Intraclass Correlations Indicating Teacher Agreement on ADHD Rating Scale Scores and Individual IRS Items by Month (Sample size, *n*)

Measures	Sep (<i>n</i> = 16)	Oct (<i>n</i> = 17)	Nov (<i>n</i> = 19)	Dec (<i>n</i> = 17)	Jan (<i>n</i> = 24)	Feb (<i>n</i> = 21)	Mar (<i>n</i> = 26)	Apr (<i>n</i> = 24)
ADHD-RS total	.70***	.40*	.26	.28	.34*	.44*	.47**	.47**
ADHD-RS inattention	.55**	.19	.07	.27	.33*	.36*	.41*	.40*
ADHD-RS Hyperactivity/impulsivity	.59**	.42*	.49 ^b	.28	.33*	.46**	.44**	.40*
IRS relationship with peers	.29	.18	.56**	.24	.34*	.21	.33*	.11
IRS relationship with teacher ^a	.38	.09	.13	.42*	.02	.15	.05	-.11
IRS academic progress	.46*	.09	.30	.40*	.45**	.36*	.35*	.47**
IRS effect on class	.30	.35	.16	.52**	.31	.30	.31	.29
IRS self-esteem ^b	.02	.22	.18	.21	.46**	.32	.16	.17
IRS overall severity ^c	.46*	.33	.15	.25	.51**	.50**	.35*	.42*

Note. ADHD-RS: Attention Deficit Hyperactive Disorder-Rating Scale; IRS: Impairment Rating Scale.

^aMarch (*n* = 21).

^bSeptember (*n* = 14), December (*n* = 14), January (*n* = 21), March (*n* = 25).

^cSeptember (*n* = 15), April (*n* = 23). * *p* < .05. ** *p* < .01. *** *p* < .001.

examined whether each teacher represents a unique context that needs to be measured independently (Kraemer et al., 2003). Answers to each of these questions provide important guidance for the assessment of school functioning and symptom manifestation of secondary school students.

Areas of Reliable Teacher Report

The reliability indices indicate that middle school teachers demonstrated the highest rates of agreement

on ratings of hyperactivity/impulsivity and academic progress. Agreement on ratings of inattention symptoms and overall impairment also were greater than .3. A review of monthly rates of agreement indicates that ratings of hyperactivity/impulsivity maintain the highest level of agreement across months. Agreement on ratings of inattention, academic progress, and overall impairment were greater than .3 for most months (>.5 during some months), but were under .15 for at least one month during the fall. These findings are consistent with previous reports that ratings of the most observable (hyperactivity/impulsivity)

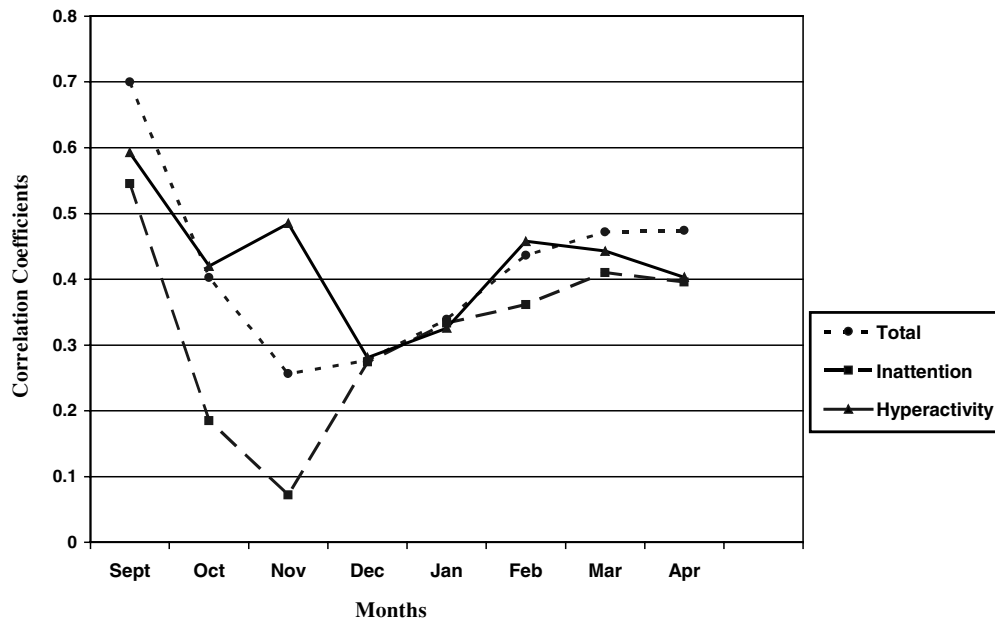


Fig. 1. Intraclass correlations between teachers on scores on the ADHD Rating Scale-IV by month.

Table III. Pearson Correlations Between Teachers' Ratings and Observational Data

Teacher ratings	On-task %	Disruptive %	<i>n</i>
ADHD-RS total	-.273***	.177*	212
ADHD-RS inattention	-.195**	.171*	212
ADHD-RS hyperactivity	-.231***	.135*	212
IRS relationship with peers	-.311***	.245***	212
IRS relationship with teacher	-.286***	.084	212
IRS academic progress	-.236***	.144*	211
IRS effect on classroom	-.340***	.179**	212
IRS self-esteem	-.282***	.134	210
IRS overall severity	-.307***	.204**	209

Note. ADHD-RS: Attention Deficit Hyperactive Disorder-Rating Scales. IRS: Impairment Rating Scales.

* $p < .05$. ** $p < .01$. *** $p \leq .001$.

behaviors tend to have the greatest inter-rater reliability (Achenbach et al., 1987). The fact that rates of agreement on measures of academic progress were relatively high is consistent with the role expectation of teachers. They spend a great deal of time measuring and grading academic work and, as a result, are likely to demonstrate the greatest assessment knowledge and expertise in this area. The inattention items on the ADHD-RS and in the DSM are clearly linked to classroom performance, since forgetting things, completing tasks, and sustaining attention are critical behaviors associated with academic performance. As a result, these are likely to be central to teachers' appraisal of a student's progress and targets of teacher attention.

The lowest rate of agreement was on the student's relationship with the teacher. This lack of agreement ($r = .06$) suggests that teacher/student relationships are specific to each teacher. Students frequently report that they get along better with some teachers than with others, and the results are consistent with that description. Teachers' rates of agreement on peer relations (.23) and self-esteem (.19) also were relatively low. Because teachers see students only in their classrooms, limited knowledge in these areas is to be expected. Furthermore, given the structure of most middle schools many students are not in classes with their friends, so teachers have very little opportunity to observe students in social situations.

The data on agreement between observation data and teacher ratings raises questions about the validity of some of the scores. For example, correlations between on-task behavior and all impairment ratings were close to .3, whereas agreement between on task behavior and ratings of ADHD symptoms were almost all less than .3. The relationship between on-task behavior and effect on the classroom was the highest whereas the relationship between on-task behavior and academic performance was

the lowest. These findings are consistent with the experience of secondary school teachers who have a wealth of data to gauge academic performance that is not a function of on-task behavior, including the completion, timely submission, and accuracy of assignments and tests. Although these data are probably not completely independent of on-task behavior, they do contribute unique variance to a teacher's appraisal of academic performance. The highest relationship between on-task behavior and student effect on classroom is intuitive since their definitions overlap. That is, students who maintain attention to tasks are unlikely to have a negative impact on the classroom, whereas those who are frequently off-task are very likely to have a negative impact on a classroom.

Ratings of a student's negative impact on a classroom are moderately associated with on-task behavior as would be expected, but on-task ratings accounted for a similar (although slightly smaller) amount of the variance in teacher ratings of peer relations, self-esteem, and relationship with teacher. Although there is reason to believe that on-task behavior might be related to impairment in all of these areas, the fact that relationships between these areas of functioning and on-task behavior are equivalent suggests another potential explanation. It might be that rating scales that ask a wide range of impairment questions force secondary school teachers to make judgments beyond their available data. In such situations they might use the data they have to make estimates of functioning in areas for which they have insufficient knowledge. This hypothesis is consistent with the finding of equivalent correlations between multiple areas of functioning and observed on-task behavior. The IRS has demonstrated specificity within domains using parent ratings (Evans, Raggi, Thompson, & Garland, 2004), but parents observe their children in multiple settings over a longer period of time than do secondary school teachers, and, as a result, parents might have more sufficient data to provide specific ratings.

Reliability Over Time

The low rates of agreement between teachers as reported in this study, as well as others (Molina et al., 1998; Simpson, 1991), could possibly be explained by a time effect. Secondary school teachers are required to get to know far more students than their elementary school counterparts, and they also have less exposure to each student. As a result, they have very little information about each student for the first few weeks of the school year, and the quantity of data increases with continued experience with the students. We hypothesized that this situation would

result in low inter-rater agreement correlations at the beginning of the year that increase over the course of the school year, and to some extent our data support this trend (see Fig. 1). A review of the data from the four ratings with the greatest inter-rater reliability (ADHD-RS Hyperactivity/Impulsivity; ADHD-RS Inattention; IRS Academic Progress; and IRS Overall) reveals that the lowest correlations on all five measures are in October, November, or December. The data from the ADHD-RS displayed in Fig. 1 reveal a trend toward improved agreement from December through April.

The results from September and October are inconsistent with our hypothesis. Some of the highest rates of agreement on the four key scales/items listed above occurred in September, declined in October and November, and then began their trend toward slow and steady improvement during the rest of the year. This initial spike in agreement might be based on very little data, but, instead, be largely a function of student reputation. As teachers develop their own data set about each child their impression of each student diverges from the reputations at unique trajectories until an accumulation of data gathered over time begins to converge as reflected in a general trend toward improved agreement. Findings for additional variables other than the four key indices do not reflect this pattern and tend to vary randomly, or agreement actually declines (e.g., relationship with teacher). It might be that working with the students in a classroom setting over time does not provide an opportunity to develop data sets about individual students that are adequate to reliably rate behavior in some of these areas (e.g., peer relations, self-esteem). Agreement on ratings of "relationships with teacher" might well be unique to individual teachers and as a result be unlikely to ever converge.

The teacher agreement data across months offer additional support for the use of the four key measures. Secondary school teachers might not be good sources of these data until January or February when rates of agreement are stable and in the moderate range, but still less than the correlation reported by Achenbach et al. (1987) whose sample consisted mostly of elementary school teachers. The mean reliability correlation for the months of February through April on the four key variables is .41, compared to only .27 for the months of October through December. These data help to explain why many researchers and clinicians have abandoned the practice of collecting teacher ratings on secondary school students. These low rates of agreement in the fall semester coupled with the non-specific relationships with on-task behavior present little reason to pursue these data, especially given that it is often difficult to collect them consistently and reliably (Pliszka et al., 2003).

Nevertheless, the evaluation of school functioning is critical to the diagnostic process and the measurement of treatment outcomes. In addition, parent reports of school functioning appear to be insufficient (Mitsis et al., 2000) for elementary school students and probably even less adequate for middle school students. An alternative to dismissing the ratings of secondary school teachers is to clarify their accumulation of knowledge about specific children, their domains of expertise, and the methods necessary to collect these data reliably and efficiently. Kraemer et al. (2003) presented a theoretical framework from which to approach this task, which involves contrasting and combining data across perspectives and contexts. Although not specifically tested in this study, implications can be drawn from these data and used to inform the testing of the model they provided.

Classrooms as Unique Contexts

Understanding whether school is one context or schools contain a variety of contexts (i.e., each classroom) has important implications for how one evaluates and implements the data integration model proposed by Kraemer et al. (2003). In their model, contexts are crossed with perspectives in a matrix that contains specific informants. Using principal-component analyses they demonstrated that data sets appropriately balanced across context and perspective produce interpretable factors including one that represents the traits or best estimate of *truth* about the person being evaluated. Kraemer et al. (2003) suggested that schools might represent a unitary context. This might be true in elementary schools, but in secondary schools each classroom appears to represent a unique context.

Clinical Implications

The data from this study suggest that classrooms are unique contexts and should be the unit of analysis when applying the Kraemer et al. (2003) model. In addition to low to moderate correlations between teacher ratings, the observational data within each month also were poorly correlated ($r = 0.25$; $p = .078$). This lack of consistency between classrooms has implications beyond the Kraemer et al. (2003) model. It also suggests that assessment data for diagnostic evaluations and treatment outcome should be collected and analyzed within classroom. Reconciling the variability that is likely to emerge when collecting ratings from various teachers might not be the appropriate goal, because if each classroom is unique, the behavior

being rated might also be unique within each classroom. The practical decisions that result from viewing each classroom as a unique context can be challenging. For example, DSM-IV requires that significant impairment due to symptoms occur across multiple settings for a diagnosis of ADHD. If classrooms are unique contexts, then could a child meet this criterion by having significant symptom-related impairment in two classrooms, but not the home? If school is considered one setting for diagnosis, then in how many classrooms must a child exhibit significant symptom related impairment for impairment at school to be considered present? These questions should be addressed in future diagnostic algorithms.

First, given the limited number of settings and amount of time that secondary school teachers interact with students it is not surprising that their agreement on ratings of peer relations, self-esteem, and relationships with teachers are very low. The inter-rater reliability data for ratings of symptoms of inattention, hyperactivity/impulsivity, academic progress, and overall impairment were in the moderate range and suggest that restricting questions to these topics might improve the quality of the data that are collected. Furthermore, additional specific questions that elicit more detailed information about these domains than what is asked in the IRS might be added to thoroughly assess areas in which teachers have expertise. For example, teachers might be able to give reliable and valid appraisals of the quantity and quality of schoolwork completed, performance on tests and quizzes, and classroom productivity. Academic progress is a critical area of functioning, especially for middle school students with ADHD, and teachers are uniquely able to provide these assessment data. Academic problems also are usually a central referral concern for parents of adolescents with ADHD (Robin, 1990; p. 463). As such, teacher ratings of academic progress are an important diagnostic and outcome data set.

Second, collecting secondary school teacher ratings prior to December of each year for anything other than hyperactivity might not be worth the effort. Rates of agreement are at their lowest and there is evidence suggesting a reputation effect at the very beginning of the school year. Teachers are just getting to know the students and many teachers respond to rating scales without consulting their data (grade book). As a result, even ratings of academic progress are likely to be based on very little information. If conducting an evaluation of an adolescent during the fall months it might be best to collect teacher ratings from the student's teachers from the previous year. Although this will not reflect current performance, previous teachers might be the best available sources for this information.

Limitations

The limitations of this study include the fact that teachers rated more than one student and that all students and teachers were sampled from the same school. Teachers rating multiple students might have reduced the variability and affected the rates of agreement. All data coming from one school might limit the generalizability of the findings. In addition, the sample size within each month appears to have been sufficient, but probably not ideal (at least within the first semester). Differences in rates of agreement over time could be attributable partly to fluctuations in sample size and because specific participants were included in some months and not others. It should also be noted that the results might have been different had teachers other than the science and math teachers been selected. These teachers were chosen because of perceived differences in rates of seatwork and group work. Teachers from classes with greater congruence between their activities might have achieved higher rates of agreement than were reported here. Finally, this was a racially and ethnically homogenous sample and teacher ratings have been reported to be influenced by racial stereotypes (Chang & Sue, 2003). Further research should attempt to address these limitations with a heterogeneous sample and determine if the findings reported here persist when these limitations are addressed.

Future Research

Numerous questions need to be addressed in future research. In particular, the evaluation model proposed by Kraemer et al. (2003) might serve as a useful tool in the assessment process of school functioning for middle school youth. Given the need to be able to accurately appraise school impairment for youth with ADHD in this age group, the application and development of assessment models is an important next step. In addition, improvements in measures are needed. As noted previously, teachers might be able to provide additional reliable data on the academic achievement, impulsivity, and inattentiveness of middle school age students beyond what is asked on current assessment instruments. Focusing item development on the factors best suited to middle school teachers will help to enhance our assessment abilities. Finally, there is a need to investigate teacher characteristics that influence ratings and find a way to integrate this source of error into the assessment process. A recent report by Gomez, Burns, Walsh, and De Moura (2003) indicated that teacher and parent characteristics are a greater influence on ratings than child characteristics in an elementary

school population. If key teacher characteristics could be assessed at the same time as the teacher rates the child, it would be possible to include them in the scoring to reduce this source of measurement error.

Conclusions

Accurate assessment of school functioning in young adolescents is critically important to the fields of mental health and education. Understanding that the scope of teachers' expertise regarding each child is more limited than their elementary school teacher counterparts and that there are important changes in their ability to report across the school year can help guide our assessment procedures. Furthermore, traditional assessment guidelines suggesting that clinicians collect data from multiple sources and look for converging themes might be confusing when applied to secondary schools unless the interpretation is guided by an understanding of the secondary school environment and the unique characteristics of each classroom.

ACKNOWLEDGMENTS

Funding for this project was provided by the Alvin V. Baird Attention and Learning Disabilities Center. The authors would like to thank the students, teachers, administrators, and parents in the Rockingham County Schools for their participation in this work. We also wish to thank Zewelanjji Serpell and Lyn Hart for this assistance with this project.

REFERENCES

- Abikoff, H. B., Jensen, P. S., Arnold, L. L. E., Hoza, B., Hechtman, L., Pollack, S., Martin, D., Alvir, J., March, J. S., Hinshaw, S., Vitiello, B., Newcorn, J., Greiner, A., Cantwell, D. P., Conners, C. K., Elliott, G., Greenhill, L. L., Kraemer, H., Pelham, W. E., Severe, J. B., Swanson, J. M., Wells, K., & Wigal, T. (2002). Observed classroom behavior of children with ADHD: Relationship to gender and comorbidity. *Journal of Abnormal Child Psychology*, *30*, 349–359.
- Abikoff, H. B., Courtney, M., Pelham, W. E., & Koplewicz, H. S. (1993). Teachers' ratings of disruptive behaviors: The influence of halo effects. *Journal of Abnormal Child Psychology*, *21*, 519–534.
- Achenbach, T. M. (1991). *Integrative guide for the 1991 CBCL/4-18, YSR, and TRF profiles*. Burlington: University of Vermont, Department of Psychiatry.
- 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.
- 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.
- American Academy of Pediatrics (2000). Clinical practice guidelines: Diagnosis and evaluation of the child with Attention-Deficit/Hyperactivity Disorder. *Pediatrics*, *105*, 1158–1170.
- American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Barkley, R. A., Edwards, G., Laneri, M., Fletcher, K., & Metevia, L. (2001). The efficacy of problem-solving communication training alone, behavior management training alone, and their combination for parent-adolescent conflict in teenagers with ADHD and ODD. *Journal of Consulting and Clinical Psychology*, *69*, 926–941.
- Bartko, J. J., & Carpenter, W. T. (1976). On the methods and theory of reliability. *Journal of Nervous and Mental Disease*, *163*, 307–317.
- Biederman, J., Faraone, S. V., Mick, E., Williamson, S., Wilens, T. E., Spencer, T. J., Weber, W., Fenton, F., Kraus, I., Pert, J., & Zallen, B. (1999). Clinical correlates of ADHD in females: Findings from a large group of girls ascertained from pediatric and psychiatric referral sources. *Journal of the American Academy of Child and Adolescent Psychiatry*, *38*, 966–975.
- Chang, D. F., & Sue, S. (2003). The effects of race and problem type on teachers' assessments of student behavior. *Journal of Consulting and Clinical Psychology*, *7*, 235–242.
- DuPaul, G. J., Power, T. J., Anastopoulos, A. D., & Reid, R. (1998). *ADHD Rating Scale—IV*. New York: The Guilford Press.
- Evans, S. W., Axelrod, J. L., & Langberg, J. M., (2004). Efficacy of a school-based treatment program for middle school youth with ADHD: Pilot data. *Behavior Modification*, *28*, 528–547.
- Evans, S. W., Langberg, J., Raggi, V., Allen, J., & Buvinger, E. (in press). Development of a school-based treatment program for middle school youth with ADHD. *Journal of Attention Disorders*.
- Evans, S. W., Raggi, V., Thompson, J., & Garland, B. (2004). *Clinical utility of the parent version of the Impairment Rating Scale for children*. Manuscript submitted for publication.
- Fabiano, G. A., & Pelham, W. E. (2002). Measuring impairment in children with attention-deficit hyperactivity disorder. *The ADHD Report*, *10*, 6–10.
- Fabiano, G. A., Pelham, W. E., Gnagy, E. M., Waschbusch, D. A., Lahey, B. B., Chronis, A. M., Onyango, A. N., Kipp, H., & Williams, A. (2004). *The reliability and validity of the Children's Impairment Rating Scale: A practical measure of impairment in children with ADHD*. Manuscript submitted for publication.
- 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.
- Greenbaum, P. E., Detric, R. F., Prange, M. E., & Friedman, R. M. (1994). Parent, teacher, and child ratings of problem behaviors of youngsters with serious emotional disturbances. *Psychological Assessment*, *6*, 141–148.
- Jensen, P. S., Rubio-Stipec, M., Canino, G., Bird, H. R., Dulcan, M. K., Schwab-Stone, M. E., & Lahey, B. B., (1999). Parent and child contributions to diagnosis of mental disorder: Are both informants always necessary? *Journal of the American Academy of Child and Adolescent Psychiatry*, *38*, 1569–1579.
- Kraemer, H. C., Measelle, J. R., Ablow, J. C., Exxes, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessments and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry*, *160*, 1566–1577.
- Lahey, B. B., Applegate, B., McBurnett, K., Biederman, J., Greenhill, L., Hynd, G. W., Barkley, R. A., Newcorn, J., Jensen, P., Richters, J., Garfinkel, B., Kerdyk, L., Frick, P. J., Ollendick, T., Perez, D., Hart, E. L., Waldman, I., & Shaffer, D. (1994). DSM-IV field trials for attention deficit hyperactivity disorder in children and adolescents. *The American Journal of Psychiatry*, *151*, 1673–1685.
- Lee, S. W., Elliott, J., & Barbour, J. D. (1994). A comparison of cross-informant behavior ratings in school-based diagnosis. *Behavioral Disorders*, *19*, 87–97.

- Merrel, K. W. (2000). Informant reports: Theory and research in using child behavior rating scales in school settings. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Theory, research, and clinical foundations* (2nd ed., pp. 233–256). New York: The Guilford Press.
- Mitsis, E. M., McKay, K. E., Schulz, K. P., Newcorn, J. H., & Halperin, J. M. (2000). Parent-teacher concordance for DSM-IV attention-deficit/hyperactivity disorder in a clinic-referred sample. *Journal of the American Academy of Child and Adolescent Psychiatry, 39*, 308–313.
- Molina, B. S. G., 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.
- Pelham, W. E., Gnagy, E., Burrows-MacLean, L., Williams, A., Fabiano, G. A., Morrisey, S. M., Chronis, A. M., Forehand, G. L., Nguyen, C. A., Hoffman, M. T., Lock, T. M., Fielbelkorn, K., Coles, E. K., Panahon, C. J., Steiner, R. L., Meichenbaum, D. L., Onyango, A. N., & Morse, G. D. (2001). Once-a-day Concerta methylphenidate versus three-times-daily methylphenidate in laboratory and natural settings. *Pediatrics, 107*, 1–15.
- Phares, V., Compas, B. E., & Howell, D. C. (1989). Perspectives on child behavior problems: Comparisons of children's self-reports with parent and teacher reports. *Psychological Assessment, 1*, 68–71.
- Pliszka, S. R., Lopez, M., Crismon, L., Toprac, M. G., Hughes, C. W., Emslie, G. J., & Boemer, C. (2003). A feasibility study of the Children's Medication Algorithm Project (CMAP) algorithm for the treatment of ADHD. *Journal of the American Academy of Child and Adolescent Psychiatry, 42*, 279–287.
- Reid, J. B., Eddy, J. M., Fetrow, R. A., & Stoolmiller, M. (1999). Description and immediate impacts of a preventive intervention for conduct problems. *American Journal of Community Psychology, 27*, 483–517.
- Reynolds, C., & Kamphaus, R. (1992). *Behavior Assessment System for Children: Manual*. Circle Pines, MN: American Guidance.
- Robin, A. L. (1990). Training families with ADHD adolescents. In R. A. Barkley (Ed.), *Attention Deficit Hyperactivity Disorder. A Handbook for Diagnosis and Treatment* (pp. 462–497). New York: Guilford Press.
- Shaffer, D., Fisher, P., Lucas, C. P., Dulcan, M. K., & Schwab-Stone, M. E. (2000). NIMH Diagnostic Interview Schedule for Children version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry, 39*, 28–38.
- Simpson, R. G. (1991). Agreement among teachers of secondary students in using the Revised Behavior Problem Checklist to identify deviant behavior. *Behavioral Disorders, 17*, 66–71.
- Wechsler, D. (1991). *Wechsler Intelligence Scale for Children-Third Edition*. New York: Psychological Corporation.