Published in final edited form as:
J Clin Child Adolesc Psychol. 2010 September ; 39(5): 650–666. doi:10.1080/15374416.2010.501287.

Incremental Validity of Test Session and Classroom Observations in a Multimethod Assessment of Attention Deficit/Hyperactivity Disorder

Stephanie H. McConaughy, Ph.D.
Department of Psychiatry, University of Vermont

Valerie S. Harder, Ph.D.
Department of Psychiatry, University of Vermont

Kevin M. Antshel, Ph.D.
Department of Psychiatry and Behavioral Sciences, SUNY-Upstate Medical University

Michael Gordon, Ph.D.
Department of Psychiatry and Behavioral Sciences, SUNY-Upstate Medical University

Ricardo Eiraldi, Ph.D.
Department of Pediatrics, University of Pennsylvania

Levent Dumenci, Ph.D.
Department of Social and Behavioral Health, Virginia Commonwealth University

Abstract

This study tested the incremental validity of behavioral observations, over and above parent and teacher reports, for assessing symptoms of Attention Deficit/Hyperactivity Disorder (ADHD) in children ages 6 to 12, using the Test Observation Form (TOF) and Direct Observation Form (DOF) from the Achenbach System of Empirically Based Assessment (ASEBA). The TOF Attention Problems and DOF Intrusive scales contributed significant unique variance, over and above parent and teacher ratings, to predicting parent and teacher ratings of hyperactivity and impulsivity and predicting categorical diagnoses of ADHD-Combined type versus Non-ADHD and ADHD-Combined type versus ADHD-Predominantly Inattentive type. The TOF Oppositional and Attention Deficit/Hyperactivity Problems scales contributed unique variance to predicting parent ratings of hyperactivity and impulsivity and the DOF Oppositional and Attention Deficit/Hyperactivity Problems scales contributed unique variance to predicting teacher ratings of hyperactivity and impulsivity.

It is now a prevailing assumption among researchers and clinicians that assessment of childhood psychiatric disorders should make use of multiple measures and informants (Johnson & Murray, 2003; Mash & Barkley, 2007). Such a "multimethod" approach is particularly important for the assessment of Attention Deficit Hyperactivity Disorder (ADHD) because symptoms and related impairments can manifest themselves in different ways across different settings and relationships. The Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition-Text Revision (DSM-IV-TR; American Psychiatric Association, 2000) provides descriptive criteria for three subtypes of ADHD: Predominantly

Correspondence regarding this article should be addressed to Stephanie H. McConaughy, Department of Psychiatry, College of Medicine, University of Vermont, One South Prospect Street, Burlington, VT 05401; Stephanie.mcconaughy@uvm.edu.

Inattentive, showing 6 of 9 specific symptoms of inattention; Predominantly Hyperactive/Impulsive, showing 6 of 9 symptoms of hyperactivity-impulsivity; and Combined, showing 6 of 9 symptoms of both inattention and hyperactivity-impulsivity. While the DSM-IV-TR does not stipulate specific procedures for assessing ADHD, the American Academy of Pediatrics (AAP, 2000) and the American Academy of Child and Adolescent Psychiatry (AACAP, 2007) have both published practice guidelines highlighting parent and teacher reports as key assessment procedures.

Many studies have shown that parent and teacher reports of attention problems, hyperactivity, and impulsivity distinguish children with ADHD from those without ADHD (for reviews, see Barkley, 2006; DuPaul & Stoner, 2003; Pelham, Fabiano, & Massetti, 2005). However, Barkley (1997) argued that relying only on parent and teacher reports in validity studies seems like circular reasoning because parent and teacher reports were the primary data used to create the ADHD diagnostic criteria in the first place. To avoid circularity, Barkley called for other assessment methods as external validators of parent and teacher reports. While numerous scientific studies have shown that deficits in behavioral inhibition and sustained attention are central to ADHD, no studies have identified a definitive neuropsychological or laboratory test for ADHD (International Consensus Statement on ADHD, 2002; National Institutes of Health, 2000; Nigg, 2006). In the absence of specific tests for ADHD, direct observations of children's behavior provide alternative methods for obtaining external validation of ADHD diagnoses. Moreover, direct observations have the advantage of being relatively “objective” when done by independent observers who have no special relationship with the child and who are “blind” to the child's clinical status. Test sessions and school classrooms are two settings that offer opportunities for independent observations of children's behavior.

Observations of Test Session Behavior

In test sessions, examiners can directly observe behavioral manifestations of ADHD symptoms while a child is engaged in cognitive or academic tasks. Several previous studies have revealed significant differences in examiners' ratings of test session behavior for children with ADHD versus typically developing control children and/or clinically referred children without ADHD (Glutting, Robins and deLancy 1997; McConaughy, Ivanova, Antshel, & Eiraldi, 2009a; Teicher, Ito, Glod & Barber, 1996; Solanto, Gilbert, Raj, Pope-Boyd, Stepak, Vail, & Newcorn, 2007) and children with the ADHD-Combined versus ADHD-Predominantly Inattentive subtypes (McConaughy et al., 2009a; Solanto et al., 2007). Other studies have found moderate correlations between test examiners' and parents' or teachers' ratings of ADHD-related problem behaviors (Glutting, Youngstrom, Oakland, & Watkins, 1996; Gordon, DiNiro, Mettelman, & Tallmadge, 1989; Willcutt, Hartung, Lahey, Loney, & Pelham, 1999).

Observations of Classroom Behavior

For many children with ADHD, observations of behavior in school classrooms can be particularly revealing, because key symptoms, such as inattentiveness, hyperactivity, and off-task behavior, may be more pronounced in school settings than in one-on-one test sessions or at home (Barkley, 2006; DuPaul & Stoner, 2003). In a review of 39 observational studies, Platzman, Stoy, Brown, Coles, Smith, and Falek (1992) concluded that, in general, classroom observations tended to be better than laboratory analog observations for distinguishing children with ADHD from comparison groups without ADHD. Volpe, DiPerna, Hintze, and Shapiro (2005) identified four classroom observation systems that showed good reliability and validity for ADHD diagnoses. Pelham et al. (2005) reviewed five observation systems for ADHD, some of which were included in the Volpe et al. (2005) review. While acknowledging the advantages of direct observations as “objective” measures of ADHD-related behaviors, Pelham et al. (2005) noted that most traditional observation systems were limited because they relied on time sampling, which failed to capture low base rate behaviors, or they were conducted in clinic analog settings, which were costly and failed to capture representative samples of behaviors that occur in natural settings. Many traditional observation systems also have the disadvantage of focusing on only a few target behaviors.

Test Session and Classroom Observations with the ASEBA Forms

Taking a different approach from traditional observation systems, McConaughy and colleagues used two forms from the Achenbach System of Empirically Based Assessment (ASEBA; Achenbach & Rescorla, 2001) to obtain examiners' ratings of test session behavior and independent observers' ratings of classroom behavior of 6-11-year-old children with and without ADHD. Using the ASEBA Test Observation Form (TOF; McConaughy & Achenbach, 2004), McConaughy et al. (2009a) found that children meeting research criteria for the ADHD-Combined type scored significantly higher than clinically referred children without ADHD and nonreferred controls on six TOF scales: Attention Problems, Oppositional, a DSM-oriented Attention Deficit/Hyperactivity Problems scale and Inattention and Hyperactivity/Impulsivity subscales, and Externalizing. The same six TOF scales also significantly discriminated between the ADHD-Combined and ADHD-Predominantly Inattentive subtypes. Examples of problem behaviors scored on the TOF scales included: argues; doesn't concentrate or pay attention for long; doesn't sit still, restless, or hyperactive; easily distracted by external stimuli; shows off, clowns or acts silly; and misbehaves or tests the limits.

Using a revised version of the ASEBA Direct Observation Form (DOF; McConaughy & Achenbach, 2009), McConaughy, Ivanova, Antshel, Eiraldi and Dumenci (2009b) found that children meeting research criteria for ADHD-Combined type scored significantly higher than clinically referred children without ADHD and controls on six DOF scales: Oppositional, Intrusive, Attention Deficit/Hyperactivity Problems, Hyperactivity/ Impulsivity, and Total Problems. The DOF Oppositional, Intrusive, and Hyperactivity-Impulsivity scales also significantly discriminated between the ADHD-Combined and ADHD-Predominantly Inattentive subtypes. The ADHD-Predominantly Inattentive subtype scored significantly higher than controls on DOF Attention Problems. In an earlier study, Skansgaard and Burns (1998) also found good discriminative validity for ADHD versus non-ADHD groups on similar scales that they created from the 1986 version of the DOF (Achenbach, 1986). Problem behaviors scored on the DOF Attention Problems and Oppositional scales were similar to those on comparable TOF scales. Examples of problem behaviors scored on the DOF Intrusive scale included: tries to get attention of staff; disturbs other children; disrupts group activities; and impatient.

Incremental Validity of Multiple Assessment Forms

Although most experts agree on the importance of multimethod assessment, few studies have examined the incremental validity of multiple methods or informants for assessment of ADHD. In simple terms, as stated by Hunsley and Meyer (2003), the concept of incremental validity addresses the question, “Does a measure add to prediction of a criterion [or dependent variable] above what can be predicted by other sources of data?” There are many different ways to test incremental validity and many different criteria (e.g., diagnosis, impairment, treatment response) against which to judge incremental validity (Haynes & Lench, 2003). With respect to ADHD, we might ask “Does a particular assessment method (e.g., behavioral observations) add to the prediction of symptom scores, or ADHD diagnoses, over and above what can be predicted by other typically used methods of assessment (e.g., parent and teacher reports)?” One approach to answering such a question, according to Hunsley and Meyers (2003), involves adding data from a new measure into hierarchical multiple regression analyses to determine the unique contribution of the new measure for predicting the dependent variable, after having entered data from one or more other measures.

Purpose of the Present Study

In the present study, we examined the incremental validity of behavioral observations for predicting parent and teacher ratings of ADHD symptoms and categorical diagnoses of ADHD-Combined type (ADHD-C) and ADHD-Predominantly Inattentive type (ADHD-IN). Following procedures of McConaughy et al. (2009a, 2009b), we obtained test examiners' ratings of 6-12-year-old children's behavior on the TOF and independent observers' ratings of their classroom behavior on the DOF. We also obtained parents' ratings of problems on the ASEBA Child Behavior Checklist for Ages 6-18 (CBCL; Achenbach & Rescorla, 2001) and teachers' ratings of problems on the ASEBA Teacher's Report Form (TRF; Achenbach & Rescorla, 2001), along with parents' and teachers' ratings of ADHD symptoms on the ADHD Rating Scale-IV Home and School versions (ADHDRS-IV; DuPaul, Power, Anastopoulos, & Reid, 1998).

We examined incremental validity of the TOF and DOF scales in three ways. First, we obtained zero-order correlations of particular TOF and DOF scales with CBCL and TRF scales and the ADHDRS-IV Inattention and Hyperactivity-Impulsivity scales. Consistent with previous research on cross-informant agreement (Achenbach, McConaughy, & Howell, 1987; Achenbach & Rescorla, 2001; Mitsis, McKay, Schulz, Newcorn, & Halperin, 2000), we expected low to moderate correlations between the TOF and DOF scales and CBCL, TRF, and ADHDRS-IV scales. We expected high correlations between CBCL, TRF, and ADHDRS-IV scales scored from ratings by the same informant. Second, we tested the unique contribution of particular TOF and DOF scales, along with the CBCL and TRF scales, for predicting continuous scores for ADHD symptoms rated by parents and teachers on the ADHDRS-IV. Third, we tested the unique contribution of particular TOF and DOF scales, along with CBCL and TRF scales, for predicting categorical research diagnoses of ADHD-C versus non-ADHD, ADHD-IN versus non-ADHD, and ADHD-C versus ADHD-IN. We expected that parent and teacher ratings on the CBCL and TRF, respectively, would contribute large percentages of variance for predicting symptom scores and categorical ADHD diagnoses. As a test of their incremental validity, we hypothesized that particular TOF and DOF scales would contribute unique variance to predicting ADHD symptom scores and categorical ADHD diagnoses and differentiating between the two ADHD subtypes.

Method

Participants

This study was part of a large federally funded research effort to test the contribution of standardized observations for assessment of ADHD. Participants in the present study were 310 children (215 boys and 95 girls) drawn from a sample of 445 children ages 6 to 12. Children were excluded from the study if they had parent-reported physical or medical problems that might interfere with test performance (e.g., seizure disorders, cerebral palsy) and/or diagnoses of mental retardation, autism, or pervasive developmental disorder. Inclusionary criteria for the sample of 310 participants were: Wechsler Intelligence Scale for Children-Fourth edition (WISC-IV; Wechsler, 2003) Full scale IQ ≥ 70, no medications for ADHD during testing, and complete data on six rating scales used as primary measures in the study (see Measures section). Forty-two cases from the total sample of 445 were excluded because they were either taking stimulant medications at the time of testing (*n* = 8) and/or they had WISC-IV FSIQ <70 or missing. Ninety-three cases were excluded because they had missing data on one or more of the six primary measures. There were no significant differences in gender distributions, mean age, or parental socioeconomic status (SES) between the final selected sample of 310 and the non-selected sample of 93 with missing data. The non-selected group scored significantly higher (mean = 54.3, *SD* = 33.7) than the selected group (mean = 44.7, *SD* = 29.9) on total problems scored on the CBCL, but there were no significant differences between the two groups on total problems scored on the TRF.

Participants were recruited from mental health providers and public and private schools in catchment areas served by outpatient clinics at three study sites: the Vermont Center for Children, Youth and Families at the University of Vermont Department of Psychiatry in Burlington, Vermont (UVM, *n* = 94); the Children's Hospital of Philadelphia in Pennsylvania (CHOP, *n* = 104); and the Child and Adolescent Psychiatry Clinic at SUNY Upstate Medical University in Syracuse, New York (SUNY, *n* = 112). The UVM clinic served a small urban and rural region, whereas the CHOP and SUNY clinics were in large urban centers. The research protocol was approved by the institutional review boards of each of the three sites. Researchers gave mental health providers and school personnel packets of consent forms and letters to parents describing the goals and procedures of the study. Parents mailed consent forms directly to the research staff, who then sent rating forms to parents and teachers and arranged appointments for testing the child at the clinic and observing the child in the school classroom. To avoid biasing selection toward concerns about ADHD per se, letters to parents described the study as an effort “to develop procedures for observing children's behavior in their classrooms and during cognitive testing.” Letters to teachers said that the child was “participating in a study of children's behavioral development.” Teachers were kept blind to referral information and test results for participants. Parents of clinically referred children and teachers were each paid $15 for their participation. The sample of 310 also included 24 children who were recruited as “normal controls.” Parents of these children were paid $50 because they presumably had less to gain from the research assessment battery than did parents of clinically referred children.

The first column of Table 1 shows demographic characteristics of the primary sample of 310 children, along with DSM-IV-TR diagnoses ascertained from the computer-generated Diagnostic Reports of the NIMH Diagnostic Interview Schedule for Children-Fourth Edition administered to parents (NIMH DISC-4; Shaffer et al., 2000; see Measures section). Children with comorbid diagnoses were counted more than once for the different diagnostic categories.

Diagnostic group assignment—From the primary sample of 310 children, 200 met research criteria for assignment to three diagnostic groups: (a) ADHD-C (*n* = 98); (b) ADHD-IN (*n* = 23), and (c) Non-ADHD (*n* = 79). (No children met criteria for ADHD-Predominantly Hyperactive/Impulsive type.) Assignment to the three diagnostic groups was based on combined parent and teacher reports of ADHD symptoms. A child was assigned to the ADHD-C group in two ways: (a) the child had a positive diagnosis of ADHD-C (314.01) on the NIMH DISC-4 Diagnostic Report (Shaffer et al., 2000; see Measures section), plus scores ≥ 80th percentile on the Inattention or Hyperactivity-Impulsivity subscales of the ADHDRS-IV-School version (DuPaul et al., 1998; see Measures section); or (b) the child had a positive diagnosis of ADHD-IN (314.00) on the NIMH DISC-4 Diagnostic Report, plus scores ≥ 80th percentile on *both* the Inattention and Hyperactivity-Impulsivity subscales of the ADHDRS-IV-School version. To be assigned to the ADHD-IN group, the child had a positive diagnosis of ADHD-IN (314.00) on the NIMH DISC-4 Diagnostic Report, plus a score ≥ 80th percentile on the Inattention subscale and a score <80th percentile on the Hyperactivity-Impulsivity subscale of the ADHDRS-IV-School version. To be assigned to the Non-ADHD group, the child had no ADHD diagnosis on the NIMH DISC-4 Diagnostic Report, and scores <80th percentile on *both* the Inattention and Hyperactivity-Impulsivity subscales of the ADHDRS-IV-School version. Children who were not assigned to one of the three diagnostic groups (*n* = 110) were those for whom parents and teachers disagreed on the research criteria for ADHD diagnoses. For example, a child might have qualified for a DSM-IV-TR diagnosis of ADHD-C based on the NIMH DISC-4, but had scores <80th percentile for teacher ratings on both the Hyperactivity-Impulsivity and Inattention scales of the ADHDRS-IV School version.

Children in the two ADHD groups were allowed to have additional DSM-IV-TR diagnoses, and children in the Non-ADHD group were allowed to have one or more DSM-IV-TR diagnoses other than ADHD. The second to fourth columns of Table 1 show the demographic characteristics and DSM-IV-TR diagnoses for each of the three diagnostic groups. It was notable that higher percentages of children in the ADHD-C group had diagnoses of Conduct Disorder (CD; 15.3%) and Oppositional Defiant Disorder (ODD; 57.1%) than those in the other two groups, and that approximately half (55.7%) of the Non-ADHD group had no DSM-IV-TR diagnosis. There were no significant age differences among the three diagnostic groups. On SES, the Non-ADHD group scored significantly higher than the ADHD-C group, *F* (2, 185) = 10.17, *p* <.001; Tukey Honestly Significant Difference (HSD) tests, *p* <.05.

Measures

ADHDRS-IV—The ADHDRS-IV (DuPaul et al., 1998) is an 18-item rating scale, with nine items assessing DSM-IV-TR symptoms of inattention and nine items assessing symptoms of hyperactivity and impulsivity. The ADHDRS-IV Home version is completed by parents and the ADHDRS-IV School version is completed by teachers. Raw scores, *T* scores, and percentiles are provided for Total Problems, Inattention, and Hyperactivity-Impulsivity, based on large stratified national samples. For the three ADHDRS-IV scales of both versions, DuPaul et al. (1998) reported internal consistency *alphas* from .86 to .96 and test-retest reliabilities, over a 4-week interval, from .78 to .90. For scores ≥ 80th percentile on the Hyperactivity-Impulsivity subscale, Power, Andrews, Eiraldi, Doherty, Ikeda, et al. (1998) reported a positive predictive probability (PPP) of .75 and negative predictive probability (NPP) of .87 for predicting ADHD-C versus controls; for scores ≥ 80th percentile on the Inattention subscale, Power et al. (1998) reported a PPP of .65 and NPP of .90 for predicting ADHD-IN versus controls.

NIMH DISC-4—The NIMH DISC-4 (Shaffer et al., 2000) is a highly structured diagnostic interview administered to parents to assess criteria for DSM-IV-TR disorders applicable to children ages 6 to 17. For this study, we administered the computer-assisted modules for ADHD, CD, ODD, anxiety disorders, mood disorders, and tic disorders. To qualify for an ADHD diagnosis on the NIMH DISC-4 Diagnostic Report, parents must report the requisite number of symptoms for one of the three ADHD subtypes, plus onset of symptoms before age 7, persistence of some symptoms for 6 or more months, and impairment from symptoms in two or more settings. Shaffer et al. (2000) reported test-retest kappas of .96 for specific phobia, .79 for ADHD, .66 for Major Depression, .65 for Generalized Anxiety Disorder, .58 for Separation Anxiety Disorder, .54 for ODD and Social Phobia, and .43 for CD.

TOF—The TOF (McConaughy & Achenbach, 2004) is a standardized rating form completed by test examiners. The TOF contains 125 items that describe children's behavior, affect, and test-taking style. During test administration, examiners record brief descriptions of the child's behavior in space provided on the TOF or on the test protocol. Immediately after the test, examiners rate the child on the 125 TOF problem items, using a 4-point scale:

0 = no occurrence; 1 = very slight or ambiguous occurrence; 2 = definite occurrence with mild to moderate intensity/frequency and less than 3 minutes total duration; 3 = definite occurrence with severe intensity, high frequency, or 3 or more minutes total duration.

The TOF problem items are scored on five empirically based syndrome scales (Withdrawn/Depressed, Language/Thought Problems, Anxious, Oppositional, and Attention Problems), a DSM-oriented Attention Deficit/Hyperactivity Problems scale with Inattention and Hyperactivity-Impulsivity subscales, Internalizing, Externalizing, and Total Problems. In addition to raw scale scores, the TOF profile provides normalized *T* scores and percentiles for each scale based on separate norms for boys and girls ages 2-5, 6-11, and 12-18. McConaughy and Achenbach (2004) reported internal consistency *alphas* ranging from .74 to .94 for the 11 TOF scales. Inter-rater reliabilities were .42 to .79 for 10 TOF scales. Test-retest reliabilities, over an average interval of 10 days, were .53 to .87 for the 11 TOF scales. For analyses in the present study, we used raw scores on the TOF Attention Problems, Oppositional, and Attention Deficit/Hyperactivity Problems scales. For these three scales, McConaughy and Achenbach (2004) reported inter-rater reliabilities of .71 to .79 and test-retest reliabilities, of .83 to .87. Criterion-related validity was demonstrated by significantly higher scores for clinically referred than nonreferred 6-11-year-old children on all TOF scales.

DOF—The DOF (McConaughy & Achenbach, 2009) is a standardized form for rating observations of children's behavior in school classrooms, at recess, and in other group settings. During a 10-minute observation period, the observer writes a narrative description of the child's behavior in space provided on the DOF. The observer also rates the child for being on-task or off-task during the last 5 seconds of each 1-minute interval. Immediately after each 10-minute observation, the observer rates the child on 89 problem items, using a 4-point scale similar to the scale for the TOF. Item 89 is open-ended for rating other problems not covered by items 1 through 88. The *0-1-2-3* item ratings are averaged across multiple 10-minute observation sessions and then summed to obtain a total raw score for each DOF problem scale. The DOF On-task score is the total number of 1-minute intervals when the child was rated as on-task, averaged across multiple 10-minute observations.

The DOF problem items are scored on five empirically based syndrome scales for classroom observations (Sluggish Cognitive Tempo, Immature/Withdrawn, Attention Problems, Intrusive, and Oppositional), a DSM-oriented Attention Deficit/Hyperactivity Problems scale with Inattention and Hyperactivity-Impulsivity subscales, and Total Problems. In addition to raw scale scores, the DOF profile provides normalized *T* scores and percentiles for each scale based on separate norms for boys and girls ages 6-11. McConaughy and Achenbach (2009) reported internal consistency *alphas* from .49 to .87 for the nine DOF problem scales. Inter-rater reliabilities were .70 to .88 for the nine DOF scales and .97 for On-task. Test-retest reliabilities were .48 to .77 for seven DOF problem scales and .42 for On-task. Criterion-related validity of the DOF was demonstrated by significantly higher scores for clinically referred than nonreferred 6-11-year-old children on all DOF scales (McConaughy & Achenbach, 2009). For the present study, we used a research edition of the DOF which included 115 problem items, but we analyzed scale scores based only the 89 problem items of the final 2009 DOF. For analyses in the present study, we used raw scores on the DOF Attention Problems, Oppositional, Intrusive, and Attention Deficit/Hyperactivity Problems scales.

CBCL and TRF—The CBCL and TRF (Achenbach & Rescorla, 2001) are standardized rating forms completed by parents and teachers, respectively. Each form includes 118 problem items, plus two open-ended problem items, that are rated on a 3-point scale: 0 = not true (as far as you know); 1 = somewhat or sometimes true; 2 = very true or often true.

CBCL ratings cover the past 6 months and TRF ratings cover the past 2 months. The CBCL and TRF profiles provide raw scores, normalized *T* scores, and percentiles for eight empirically based syndrome scales, six DSM-oriented scales, Internalizing, Externalizing, and Total Problems. The CBCL profile also provides scores for competence scales, while the TRF profile provides scores for academic performance and adaptive functioning. The CBCL and TRF are normed separately for boys and girls ages 6-11 and 12-18. Achenbach and Rescorla (2001) reported internal consistency *alphas* of .72 to .97 for the CBCL and .72 to .95 for the TRF problem scales. For test-retest reliabilities, Achenbach and Rescorla (2001) reported mean *r*s of .90 averaged separately across the CBCL and TRF empirically based scales, a mean *r* of .88 for the CBCL DSM-oriented scales, and a mean *r* of .85 for the TRF DSM-oriented scales. All CBCL and TRF problem scales significantly discriminated between matched samples of clinically referred and non-referred children. For analyses in the present study, we used raw scores on the CBCL and TRF Attention Problems and DSM-oriented Attention Deficit/Hyperactivity Problems scales.

Procedures

Test session observations—Test examiners were advanced psychology or school psychology graduate students trained in test administration. Examiners administered three tests to each child: the Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV; Wechsler, 2003); the Wechsler Individual Achievement Test-Second Edition (WIAT-II; Wechsler, 2002); and the Gordon Diagnostic System Model III (GDS; Gordon, 1989). The GDS is a computerized continuous performance test that measures impulse control, sustained attention and distractibility. The order of the tests was counter balanced across children. Most children were tested in one session, lasting approximately 3 hours. Test examiners had the option of dividing testing into two sessions on different days for the few children who needed shorter sessions. Researchers trained test examiners in the rating procedures described in the TOF manual (McConaughy & Achenbach, 2004). Test examiners were also provided written guidelines and exemplars for rating the 125 TOF problem items. Examiners completed a separate TOF immediately after administering (and before scoring) each of the three tests (WISC-IV, WIAT-II, and GDS). Children had a 15-minute break with a sugar-free snack between tests. Prior to testing, examiners were kept blind to all information about the child, including referral concerns, scores on parent and teacher rating scales, and the child's diagnostic group assignment for this study.

To assess inter-rater reliabilities for test observations in the present study, eight trained test examiners rated a total of 47 videotapes of WISC-IV test sessions that were conducted by another trained examiner. Inter-rater reliabilities (p <.001) for the three TOF scales used in this study were .69 for Attention Problems, .60 for Oppositional, and .77 for Attention Deficit/Hyperactivity Problems. In addition, four examiners rated a total of 35 videotapes of their own test sessions, with an average interval of 9.9 months between Time 1 and Time 2 ratings. Inter-rater reliabilities for ratings by the same examiner (p <.001) were .80 for Attention Problems, .53 for Oppositional, and .71 for Attention Deficit/Hyperactivity Problems (Turkoglu, 2009).

Classroom observations—Across the three study sites, there were 24 classroom observers, including undergraduate psychology students and post-graduates with Bachelor's or Master's degrees. Researchers trained observers in the recording and rating procedures described in the DOF manual (McConaughy & Achenbach, 2009). After training, observers used the DOF to obtain three to four 10-minute observations of each child participant. (Ten participants had three observations and 300 had four observations.) Observations were conducted on two separate days with two observations in the morning and two observations in the afternoon. The median interval between the first and last observation was 1 day (25th quartile = 1 day; 75th quartile = 4 days). Observers were instructed to conduct their observations only during academic activities (e.g., reading, math, science, social studies) and not during free time. The type of academic activity varied across students and sometimes changed within a 10-minute observation period. Observers followed the procedures described in the Measures section for recording observations of the child's behavior. Observers were also provided written guidelines and exemplars for rating the 89 DOF problem items, as described in the DOF manual (McConaughy & Achenbach, 2009). Observers were blind to all information about the child, including referral concerns, test scores, parent and teacher rating scale scores, and the child's diagnostic group assignment for this study. Observers were also instructed not to discuss their observations with teachers.

To assess inter-rater reliabilities for this study, 12 pairs of trained observers rated 1 to 4 10-minute classroom observations of 212 randomly selected children (112 boys, 100 girls) in elementary schools near the three study sites. Observer pairs rated 14 to 24 randomly selected children. Observers were instructed not to discuss their ratings with each other until after all reliability data were collected. Averaged across the 12 pairs of observers, inter-rater reliabilities for the four DOF scales used in this study were .72 for Attention Problems, .71 for Oppositional, .78 for Intrusive, and .80 for Attention Deficit/Hyperactivity Problems.

Statistical Analyses

As a first step in our analyses, we obtained Pearson correlations between the TOF, DOF, CBCL, TRF, and ADHDRS-IV scales. As a second step, we conducted hierarchical multiple regressions to test the unique contributions of specific TOF and DOF scales, over and above CBCL and TRF scales, for predicting continuous symptom scores on the ADHDRS-IV. In the hierarchical multiple regressions, we determined the order of entry of predictors in forced step-wise procedures, as recommended by Hunsley and Meyer (2003). Because there was high collinearity among the three TOF scales and among the four DOF scales used in this study, the TOF and DOF scales were entered as predictors in separate models. Dependent variables were raw scores on the Inattention and Hyperactivity-Impulsivity scales of the ADHDRS-IV Home and ADHDRS-IV School versions. Preliminary hierarchical multiple regressions showed no significant associations of demographic variables (gender, age, SES), entered in Step 1. Therefore, the demographic variables were eliminated from final models. Seven hierarchical multiple regressions were conducted for each of the four dependent variables (a total of 28 hierarchical multiple regressions as primary analyses). In 5 of the 7 models, CBCL and TRF Attention Problems were entered together in Step 1, and one of the following TOF or DOF scales was entered in Step 2: TOF or DOF Attention Problems, TOF or DOF Oppositional, and DOF Intrusive. In 2 of the 7 models, CBCL and TRF Attention Deficit/Hyperactivity Problems were entered together in Step 1, and TOF or DOF Attention Deficit/Hyperactivity Problems was entered in Step 2. Unique variance accounted for by the TOF and DOF scales was measured by the change in R^2 (ΔR^2) between Step 1 and Step 2. Unique variance accounted for by each predictor was also measured by the semi-partial r^2 at Step 1 and Step 2. (For the TOF and DOF scales, the semi-partial r^2 at Step 2 was equal to ΔR^2).

We examined Tolerance and Variance Inflation Factors (VIF) as tests of multicollinearity among predictors in the regression analyses. All Tolerance values were well above .20 and all VIFs were well below 4.0, which are commonly accepted cut-off criteria. We also examined the effects of potential outliers in regression analyses, using Cook's D >1.0 and Studentized residuals >3.0 as indicators of possible outlier problems. In all 28 regressions, Cook's D was well below the general guideline criterion of 1.0 (Howell, 2010). Only one to three outliers were identified by Studentized residuals. After removing outliers, there were no changes in patterns of effects originally significant at p <.01 in any model. Slight changes occurred in patterns of effects originally significant at *p* <.05 in four models, which are marked in tables in Results.

Primary hierarchical multiple regressions used TOF scale scores based on observations during the WISC-IV, for which there were no missing values (*N*= 310). Secondary hierarchical multiple regressions were also performed substituting TOF scores from observations during the WIAT-II (*N* = 309) or the GDS (*N* = 302), using listwise deletion for missing values. We analyzed TOF scores for the WISC-IV, WIAT-II, and GDS separately because researchers and clinicians may not choose to administer all three tests to every child. Moreover, repeated measures ANOVAs showed significant mean differences between TOF scale scores based on the WISC-IV versus TOF scores based on the GDS, for Attention Problems, *F* (1, 301) = 57.34, *p* <.001, and Attention Deficit/Hyperactivity Problems, *F* (1, 301) = 28.02, *p* <.001. There were no significant differences in TOF Attention Problems scores based on the WISC-IV versus WIAT-II or for TOF Oppositional scores based on any of the three tests.

As a third test of incremental validity, we conducted multinomial logistic regressions to test the relative contributions of the TOF and DOF scales, along with CBCL and TRF scales, for predicting categorical diagnoses of ADHD-C versus Non-ADHD, ADHD-IN versus Non-ADHD, and ADHD-C versus ADHD-IN. These analyses were carried out on the subsample of 200 children who met research criteria for DSM-IV-TR diagnoses of ADHD-C (*n* = 98), ADHD-IN (*n* =23), and Non-ADHD (*n* =79), as described in the Method section. The combinations of predictors for the multinomial logistic regressions were the same as those used in the final models for hierarchical multiple regressions.

All of our analyses were conducted with PSAW SPSS 17.0 (2009). To correct for experiment-wise error rate, we determined the number of significant effects (significant Beta weights at Step 2) at *p* <.05 that might be expected by chance, using a *p* <.05 protection level, based on graphs and procedures described by Sakoda, Cohen, and Beall (1954).[1]

Results

Descriptive Statistics

Tables 2 and 3 report means, standard deviations and intercorrelations among the CBCL, TRF, TOF-WISC-IV, DOF and the ADHDRS-IV scales. Correlations were low to moderate between the observational measures paired with parent and teacher measures and higher between pairings of parent measures and pairings of teacher measures.

Hierarchical Multiple Regressions

Table 4 shows results of hierarchical multiple regressions for the three models using one of three TOF scales, along with CBCL and TRF scales, as predictors of Hyperactivity-Impulsivity scores on the ADHDRS-IV Home and School versions. Total variance accounted for by all predictors at Step 2 ranged from 48 to 64% for parent-rated Hyperactivity-Impulsivity and 54 to 72% for teacher-rated Hyperactivity-Impulsivity. As expected, CBCL Attention Problems and CBCL Attention Deficit/Hyperactivity Problems contributed most to predicting parent-rated Hyperactivity-Impulsivity, accounting for 30 to

[1] Across the three models using the CBCL, TRF, and TOF scales as predictors (9 statistical tests at Step 2), two effects (significant Beta weights) at *p* <.05 might be due to chance, using a *p* <.05 protection level. Across the four models using the CBCL, TRF, and DOF scales as predictors (12 statistical tests at Step 2), 2 to 3 effects (significant Beta weights) at *p* <.05 might be due to chance, using a *p* <.05 protection level (Sakoda, Cohen, & Beall, 1954). Possible chance effects are marked in tables for sets of hierarchical multiple regressions for each dependent variable and each set of multinomial logistic regressions. Effects at *p* <.01 and *p* <.001 were not considered possible chance effects.

43% of variance in Step 2 (see semi-partial r^2 for CBCL scales). Also as expected, TRF Attention Problems and TRF Attention Deficit/Hyperactivity Problems contributed most to predicting teacher-rated Hyperactivity-Impulsivity, accounting for 44 to 54% of variance in Step 2 (see semi-partial r^2 for TRF scales). TOF Attention Problems based on the WISC-IV (Model 1a) was a significant independent predictor of parent- and teacher-rated Hyperactivity-Impulsivity, accounting for 1 to 3% of variance, over and above CBCL and TRF Attention Problems (see ΔR^2 and semi-partial r^2 for TOF). TOF Oppositional based on the WISC-IV (Model 1b) was a significant independent predictor of parent-rated Hyperactivity-Impulsivity, accounting for 2% of variance, over and above the CBCL and TRF Attention Problems. TOF Attention Deficit/Hyperactivity Problems based on the WISC-IV (Model 1c) was a significant independent predictor of parent-rated Hyperactivity-Impulsivity, accounting for 1% of variance, over and above the CBCL and TRF Attention Deficit/Hyperactivity Problems, but this could have been a chance effect when corrected for the number of analyses (Sakoda et al., 1954).

Secondary hierarchical multiple regressions showed similar contributions of the CBCL, TRF and TOF scales for predicting parent- and teacher-rated Hyperactivity-Impulsivity when TOF scores based on observations during the WIAT-II or the GDS were substituted for TOF scores based on the WISC-IV. By contrast, hierarchical multiple regressions showed no significant unique contributions of any of the three TOF scales for predicting parent- or teacher-rated Inattention scores on the ADHDRS-IV Home and School versions.

Table 5 shows results of hierarchical multiple regressions for the four models using one of four DOF scales, along with CBCL and TRF scales, as predictors of Hyperactivity-Impulsivity scores on the ADHDRS-IV Home and School versions. Total variance accounted for by all predictors in Step 2 ranged from 46 to 63% for parent-rated Hyperactivity-Impulsivity and 54 to 72% for teacher-rated Hyperactivity-Impulsivity. Again as expected, CBCL Attention Problems and CBCL Attention Deficit/Hyperactivity Problems contributed most to predicting parent-rated Hyperactivity-Impulsivity, accounting for 34 to 47% of variance in Step 2. TRF Attention Problems and TRF Attention Deficit/Hyperactivity Problems contributed most to predicting teacher-rated Hyperactivity-Impulsivity, accounting for 38 to 50% of variance in Step 2. DOF Oppositional (Model 2b) was a significant independent predictor of teacher-rated Hyperactivity-Impulsivity, accounting for 2% of variance, over and above CBCL and TRF Attention Problems. DOF Intrusive (Model 2c) was a significant independent predictor of parent- and teacher-rated Hyperactivity-Impulsivity, accounting for 2 to 6% of variance, over and above CBCL and TRF Attention Problems. DOF Attention Deficit/Hyperactivity Problems (Model 2d) was a significant independent predictor of teacher-rated Hyperactivity-Impulsivity, but this could have been a chance effect (Sakoda, et al., 1954) and the percent of variance accounted for was <1%.

Hierarchical multiple regressions for the predicting Inattention scores on the ADHDRS-IV Home and School versions showed only one significant effect for the DOF scales: DOF Attention Deficit/Hyperactivity Problems made a significant unique contribution to predicting parent-rated Inattention (Beta = -.08, p <.05, semi-partial r^2 = .01), but this could have been a chance effect (Sakoda et al., 1954).

Multinomial Logistic Regressions

Table 6 shows results from multinomial logistic regressions for the three models using one of three TOF scales, along with CBCL and TRF scales, as predictors of categorical diagnostic classifications of ADHD-C versus Non-ADHD and ADHD-C versus ADHD-IN. The combination of CBCL, TRF and TOF scales across the three models accounted for 67 to 72% of variance. CBCL and TRF Attention Problems were significant predictors of ADHD-

C versus Non-ADHD, but not ADHD-C versus ADHD-IN. TOF Attention Problems-WISC-IV was a significant predictor of both ADHD-C versus Non-ADHD and ADHD-C versus ADHD-IN (Model 3a), but these could have been chance effects when corrected for the number of analyses. CBCL and TRF Attention Deficit/Hyperactivity Problems were significant predictors of both ADHD-C versus Non-ADHD and ADHD-C versus ADHD-IN, but TOF Attention Deficit/Hyperactivity Problems was not. Multinomial logistic regressions showed no significant effects of any TOF scale for predicting ADHD-IN versus Non-ADHD.

Table 7 shows results from multinomial logistic regressions for the four models using one of four DOF scales, along with CBCL and TRF scales, as predictors of ADHD-C versus Non-ADHD and ADHD-C versus ADHD-IN. The combination of CBCL, TRF and DOF scales across the four models accounted for 67 to 73% of variance. CBCL and TRF Attention Problems were significant predictors of ADHD-C versus Non-ADHD, but not ADHD-C versus ADHD-IN. Relatively robust effects were found for DOF Intrusive (Model 4c), which was a significant predictor of both ADHD-C versus Non-ADHD and ADHD-C versus ADHD-IN, with odds ratios larger than those for CBCL and TRF Attention Problems. DOF Oppositional (Model 4b) was a significant predictor of ADHD-C versus ADHD-IN, and DOF Attention Deficit/Hyperactivity Problems (Model 4d) was a significant predictor of ADHD-C versus Non-ADHD, but these could have been chance effects. Multinomial logistic regressions showed no significant effects of any DOF scale for predicting ADHD-IN versus Non-ADHD.

Discussion

In this study, we tested the incremental validity of observational methods, over and above parent and teacher reports, in multimethod assessment of ADHD. We considered two contexts in which children's behavior can be readily observed in multimethod assessments: test sessions and school classrooms. We used two standardized ASEBA rating forms, the TOF and DOF, as our observational measures, along with the CBCL and TRF for parent and teacher reports. These four ASEBA rating forms have the advantage of sharing many comparable problem items and similar problem scales.

As expected, we found low to moderate correlations between the TOF, DOF, CBCL, and TRF scales, and between the TOF, DOF, and ADHDRS-IV scales (Tables 2 and 3, *r* = .07 to .39). These findings are consistent with previous research showing moderate agreement between different types of informants (Achenbach et al., 1987;Achenbach & Rescorla, 2001;Mitsis et al., 2000). The low to moderate correlations also showed that the TOF and DOF scales were not redundant with the other two predictors (CBCL and TRF scales), or our dependent measures (ADHDRS-IV scales), which is a prerequisite for demonstrating incremental validity (Haynes & Lench, 2003). By contrast, correlations were much higher for pairings of CBCL and TRF scales with ADHDRS-IV subscales from the same informants (Tables 2 and 3, *r* = .66 to .83), consistent with research showing higher agreement between similar informants than different informants (Achenbach et al., 1987;Mitsis et al., 2000).

In hierarchical multiple regressions, we entered the CBCL and TRF Attention Problems or Attention Deficit/Hyperactivity Problems scales in Step 1 because parent and teacher ratings represent a common and efficient method for assessing ADHD. After entering the observational measures in Step 2 in hierarchical multiple regressions, we found that particular TOF and DOF scales contributed unique variance to predicting parent and teacher reports of ADHD symptoms, over and above the contributions of the CBCL and TRF scales. Specifically, the TOF Attention Problems and DOF Intrusive scales made unique contributions to predicting both parent- and teacher-rated hyperactivity-impulsivity, accounting for 1 to 6% of variance. The TOF Oppositional scale made a unique contribution to predicting parent-rated hyperactivity-impulsivity and the DOF Oppositional scale made a unique contribution to predicting teacher-rated hyperactivity-impulsivity, accounting for 2% of variance. The TOF and DOF Attention Deficit/Hyperactivity Problems scales also contributed unique variance to predicting parent- and teacher-rated hyperactivity-impulsivity, respectively, but these could have been chance effects. These findings support the incremental validity of observations of test session and classroom behavior in multimethod assessment of ADHD. Willcutt et al. (1999) also reported that ratings by non-clinician test examiners contributed significant variance, over and above parent and teacher ratings, to predicting functional impairment in preschool children with ADHD.

Our findings for the TOF and DOF Oppositional scales, and DOF Intrusive scale, are consistent with relatively high rates of parent-reported ODD in our sample (35%), particularly for children with ADHD-C (57%). Many other studies have also shown high comorbidity between ODD and ADHD-C (e.g., Jensen, Hinshaw, Kraemer, Lenora, Newcorn, Abikoff, et al, 2001). Such findings underscore the importance of directly observing children's disruptive behaviors in test sessions and school classrooms in addition to observing problems indicative of ADHD symptoms per se. It is also important to consider the possibility of “halo effects” of ODD behaviors on test examiners' and classroom observers' ratings of hyperactivity and impulsivity, as shown in an analog study by Abikoff, Courtney, Pelham and Koplewicz (1993). Abikoff and colleagues had elementary teachers observe 10-minute video tapes of child actors depicting deviant levels of “pure ADHD” behaviors, deviant levels of “pure ODD” behaviors but normal levels of ADHD behaviors, and normal levels of both types of behaviors. They found that the teacher observers rated the child displaying ODD behaviors significantly higher on hyperactivity and impulsivity than the normal child, suggesting a bias toward inflated hyperactivity-impulsivity ratings in the presence of ODD behaviors. However, unlike the Abikoff et al. study, many children in our sample qualified for diagnoses of both ADHD and ODD. So while our test examiners and classroom observers may have rated hyperactivity and impulsivity higher for children with ODD, it is likely that this was because many of these children actually exhibited both ADHD and ODD behaviors.

Results from the multinomial logistic regressions provided further support for the incremental validity of the TOF Attention Problems, DOF Intrusive and DOF Oppositional scales. It was notable that the DOF Intrusive scale contributed most to predicting categorical diagnostic classifications of ADHD-C versus Non-ADHD and ADHD-C versus ADHD-IN, with odds ratios larger than those for CBCL and TRF Attention Problems. Interestingly, CBCL and TRF Attention Problems made no significant contributions to classification of ADHD-C versus ADHD-IN. However, results for classification of the ADHD subtypes should be viewed with caution because of the disproportionate sample sizes.

While our results supported the incremental validity of several TOF and DOF scales for predicting hyperactivity-impulsivity, the percent of unique variance accounted for by the TOF and DOF scales was small, according to Cohen's criteria (1988). This was not surprising, since we relied on parent and teacher ratings of ADHDRS-IV symptoms for our dependent measures, and at the same time used parent and teacher ratings on the CBCL and TRF scales as predictors. As Pelham et al. (2005) pointed out, psychometricians have long acknowledged the challenges presented by source and method variance in studies of ADHD. However, as noted by Barkley (1997), such circularity in assessment cannot be avoided without an independent gold standard criterion for ADHD.

Contrary to our expectations, the TOF and DOF scales did not contribute significant variance to predicting parent or teacher ratings of inattention symptoms on the ADHDRS-IV. This was likely due to the fact that parent and teacher ratings on the CBCL and TRF Attention Problems and Attention Deficit/Hyperactivity Problems scales accounted for so much variance (54 to 57%) in predicting inattention that there was little variance left to be accounted for by the observational measures. Similarly, in the multinomial logistic regressions, only the CBCL and TRF scales predicted categorical diagnostic classifications of ADHD-IN versus NON-ADHD. These findings suggest that ratings of children's behavior by test examiners and classroom observers add little to no unique information, over and above parent and teacher ratings, in multimethod assessment of inattention symptoms. This may be because inattention symptoms are more difficult to observe than hyperactivity and impulsivity and so are better reported by teachers and parents who have more experience with the child. It is also possible that different observational approaches, such as partial interval coding or event coding, obtained over longer durations, would be more sensitive to detecting inattention in multimethod assessment.

Limitations

There are several limitations to our study. First, the *s*ample size was relatively small for ADHD-IN (*n* = 23) compared to the other two groups. The small sample size could have reduced power for finding contributions of the TOF and DOF scales for predicting ADHD-IN versus Non-ADHD or ADHD-IN versus ADHD-C in the multinomial logistic regressions. A second limitation was that our sample included only 6-12-year-old children, so the results may not generalize to adolescents. A third limitation was that classroom observers and test examiners may have developed some hypotheses about the children that could have affected their ratings on the TOF and DOF. To minimize rater bias, classroom observers and test examiners were kept blind to all clinical information about the children and they were given detailed behavioral descriptors as guidelines for scoring the TOF and DOF problem items. Fourth, there were some limitations in assessment of the reliability of our observational measures. During our training procedures, we obtained moderate to high inter-rater reliabilities for the four DOF scales (*r* = .72 to .80) across our 24 observers, based on observations of anonymous non-participant children. However, we did not obtain additional inter-rater reliabilities for subsamples of participants in this particular study, as done in some other observational studies. For test examiners in this study, we obtained modest inter-rater reliabilities for the TOF Oppositional scale (*r* = .53 and .60). Although McConaughy and Achenbach (2004) reported higher inter-rater reliability for TOF Oppositional (*r* = .79), our lower reliabilities might have placed some limits on correlations between this particular scale and our dependent measures. A final limitation was that the present study focused only on ADHD symptoms and diagnoses. Additional research is under way to test incremental validity of test session and classroom observations for predicting social and academic impairments associated with ADHD.

Implications for Research, Policy, and Practice

Considering the modest agreement usually found between parents and teachers (Achenbach et al., 1987: Mitsis et al., 2000), independent observations of children's behavior provide an additional venue for validating ADHD diagnoses. As Pelham et al. (2005) pointed out, cross-informant agreement between parent and teacher ratings can be expected to be low because raters have different tolerance levels and different interpretations of children's behaviors and children often behave differently in different situations, such as home versus school. In addition, the low correlations between the TOF and DOF scales found in this study (Table 2) suggest that direct observations in test sessions and school classrooms capture different aspects of children's behavior, making it important to observe children in multiple contexts.

Although specific tests are not required for ADHD diagnoses, clinic-based and school-based practitioners often administer intelligence and achievement tests as components of ADHD assessments (Barkley, 2006; Demaray, Schaefer, & Delong, 2003; DuPaul & Stoner, 2003). Moreover, both the AAP and AACAP encourage standardized testing whenever academic underachievement or cognitive functioning is a concern. The TOF provides a standardized and efficient method for recording and quantifying observations of a wide array of test session behaviors. Once the rating guidelines become familiar, an examiner should be able to complete TOF ratings in about 10 minutes. The TOF scales can then be scored via computer by the examiner or support staff. These considerations support the clinical utility and efficiency of the TOF.

Observations of children's classroom behavior require additional time beyond time invested in other assessment procedures. As Pelham et al. (2005) noted, classroom observations may be impractical for clinic-based practitioners because of additional costs, need for trained observers, and need for multiple observations. However, school psychologists often conduct classroom observations to assess behavior problems, including ADHD (Demaray et. al., 2003; Shapiro & Heick, 2004). Paraprofessional staff can also be trained to use the DOF for school-based observations. The DOF provides a standardized method for recording and quantifying observations. The DOF requires approximately 10 minutes for each observation and an additional 5 minutes per observation to rate the DOF items. A computer program can then be used to score multiple observations on the DOF scales. These considerations support the clinical utility and efficiency of the DOF, particularly for school psychologists and consulting clinical psychologists. In addition, clinic-based practitioners can collaborate with school-based practitioners to obtain observations for ADHD assessments, as done routinely by one of the authors of this article (McConaughy).

Finally, keeping in mind the modest unique contributions of behavioral observations for ADHD assessment, the TOF and DOF might best be used in research and clinical practice in several ways. First, the DOF can be used to screen for problem behaviors in schools that warrant further comprehensive assessment. Second, in comprehensive assessments, high scores on relevant DOF and TOF scales can be used to corroborate parent and teacher reports. These observational measures might be particularly useful when parents and teachers disagree in their reports of ADHD symptoms. Practitioners should pay special attention to high scores on those scales that showed discriminative validity for ADHD in previous studies (McConaughy et al., 2009a, 2009b) and incremental validity for ADHD in the present study. Third, the TOF and DOF can be used to directly assess oppositional and intrusive behaviors that often co-occur with ADHD symptoms. Our findings for the TOF and DOF, along with those of McConaughy et al. (2009a, 2009b), are consistent with many other studies showing that children with ADHD often display a multiplicity of problems in addition to ADHD symptoms (for reviews, see Barkley, 2006; DuPaul & Stoner, 2003). Fourth, the DOF can be used, along with parent and teacher rating scales, to evaluate and monitor effects of school-based interventions for ADHD. Supporting the practical utility of the DOF for such formative assessments, Volpe, McConaughy and Hintze (2009) reported that the DOF Oppositional, Intrusive, and Attention Deficit/Hyperactivity Problems scales, as well as the DOF Hyperactivity-Impulsivity subscale, showed strong generalizability and dependability over multiple 10-minute observations and required fewer observations than other DOF scales to reach acceptable reliability. However, future research is needed to determine whether the DOF is sensitive to small behavioral changes that may occur in response to interventions.

Acknowledgments

Preparation of this article was supported in part by Grant R01 HD40220 from the National Institute of Child Health and Human Development to Stephanie McConaughy at the University of Vermont and by the University of Vermont Research Center for Children, Youth, and Families. Statements do not represent the position or policy of these agencies and no official endorsement by them should be inferred.

References

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

Achenbach, TM. The Direct Observation Form of the Child Behavior Checklist (rev ed). Burlington, VT: University of Vermont, Department of Psychiatry; 1986.

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

Achenbach, TM.; Rescorla, LA. Manual for the ASEBA School-Age Forms & Profiles. Burlington, VT: University of Vermont, Research Center for Children, Youth, and Families; 2001.

American Academy of Child and Adolescent Psychiatry. Practice parameter for assessment and treatment of children and adolescents with Attention Deficit/Hyperactivity Disorder. Journal of the American Academy of Child and Adolescent Psychiatry 2007;46:894–921. [PubMed: 17581453]

American Academy of Pediatrics. Diagnosis and evaluation of the child with Attention-Deficit/Hyperactivity Disorder. Pediatrics 2000;105:1158–1170. [PubMed: 10836893]

American Psychiatric Association. Diagnostic and statistical manual of mental disorders: Fourth edition: Text revision. Washington, D.C.: Author; 2000.

Barkley, RA. ADHD and the nature of self-control. New York: Guilford Press; 1997.

Barkley, RA. Attention deficit hyperactivity disorder: A handbook for diagnosis and treatment. 3rd. New York: Guilford Press; 2006.

Cohen, J. Statistical power analysis for the behavioral sciences. 2nd. New York: Academic Press; 1988.

Demaray MK, Schaefer K, Delong LK. Attention deficit/hyperactivity disorder (ADHD): A national survey of training and assessment practices in the schools. Psychology in the Schools 2003;40:583–597.

DuPaul, GJ.; Power, T.; Anastopoulos, AD.; Reid, R. Manual for the ADHD Rating Scale-IV. New York: Guilford Press; 1998.

DuPaul, GJ.; Stoner, G. ADHD in the schools. 2nd. New York: Guilford Press; 2003.

Glutting JJ, Robins PM, de Lancy E. Discriminant validity of test observations for children with Attention Deficit Hyperactivity Disorder. Journal of School Psychology 1997;35:391–401.

Glutting JJ, Youngstrom EA, Oakland T, Watkins MW. Situational specificity and generality of test behaviors for samples of normal and referred children. School Psychology Review 1996;25:94–107.

Gordon, M. Gordon Diagnostic System Model III. Dewitt, NY: Gordon Systems; 1998.

Gordon M, DiNiro D, Mettelman BB, Tallmadge J. Observations of test behavior, quantitative scores, and teacher ratings. Journal of Psychoeducational Assessment 1989;7:141–147.

Haynes SN, Lench HC. Incremental validity of new clinical assessment measures. Psychological Assessment 2003;15:456–466. [PubMed: 14692842]

Hollingshead, AB. Four factor index of social status. New Haven, CT: Yale University, Department of Sociology; 1975. Unpublished paper

Howell, David, C. Statistical methods for psychology. 7th. Belmont, CA: Cengage Wadsworth; 2010.

Hunsley J, Meyer GJ. The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. Psychological Assessment 2003;15:446–455. [PubMed: 14692841]

International Consensus Statement on ADHD. Clinical Child and Family Psychology Review 2002;5:89–111. [PubMed: 12093014]

Jensen PT, Hinshaw SP, Kraemer HC, Lenora N, Newcorn JH, Abikoff HB, et al. ADHD comorbidity findings from the MTA Study: Comparing comorbid subtypes. Journal of the American Academy of Child and Adolescent Psychiatry 2001;40:147–158. [PubMed: 11211363]

Johnson C, Murray C. Incremental validity in the psychological assessment of children and adolescents. Psychological Assessment 2003;15:496–507. [PubMed: 14692845]

Mash, EJ.; Barkley, RA. Assessment of childhood disorders. New York: Guilford Press; 2007.

McConaughy, SH.; Achenbach, TM. Manual for the Test Observation Form for Ages 2-18. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families; 2004.

McConaughy, SH.; Achenbach, TM. Manual for the ASEBA Direct Observation Form. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families; 2009.

McConaughy SH, Ivanova M, Antshel K, Eiraldi RB. Standardized observational assessment of Attention Deficit/Hyperactivity Disorder Combined and Predominantly Inattentive Subtypes: I. Test session observations. School Psychology Review 2009a;38:45–66.

McConaughy SH, Ivanova M, Antshel K, Eiraldi RB, Dumenci L. Standardized observational assessment of Attention Deficit/Hyperactivity Disorder Combined and Predominantly Inattentive Subtypes: II. Classroom observations. School Psychology Review 2009b;39:362–381.

Mitsis EM, McKay KE, Schulz KP, Newcorn JH, Halperin JM. 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 2000;39:308–313. [PubMed: 10714050]

National Institutes of Health. Consensus development conference statement: Diagnosis and treatment of Attention-Deficit/Hyperactivity Disorder (ADHD). Journal of the American Academy of Child and Adolescent Psychiatry 2000;39:182–193. [PubMed: 10673829]

Nigg, JT. What causes ADHD?. New York: Guilford Press; 2006.

Pelham WE, Fabiano GA, Massetti GM. Evidence-based assessment of Attention Deficit Hyperactivity Disorder in children and adolescents. Journal of Clinical Child and Adolescent Psychology 2005;34:449–476. [PubMed: 16026214]

Platzman KA, Stoy MR, Brown RT, Coles C, Smith IE, Falek A. Review of observational methods in Attention Deficit Hyperactivity Disorder (ADHD): Implications for diagnosis. School Psychology Quarterly 1992;7:155–177.

Power TJ, Andrews TJ, Eiraldi RB, Doherty BJ, Ikeda MJ, DuPaul GJ, Landau S. Evaluating Attention Deficit Hyperactivity Disorder using multiple informants: The incremental utility of combining teacher with parent reports. Psychological Assessment 1998;10:250–260.

PSAW SPSS. SPSS Base 17.0 User's Guide. Chicago, IL: PSAW SPSS; 2009.

Sakoda JM, Cohen BH, Beall G. Test of significance for a series of statistical tests. Psychological Bulletin 1954;51:172–175. [PubMed: 13155709]

Skansgaard EP, Burns GL. Comparison of DSM-IV ADHD combined and predominantly inattention types: Correspondence between teacher ratings and direct observations of inattentive, hyperactivity/impulsivity, slow cognitive tempo, oppositional defiant, and overt conduct disorder symptoms. Child & Family Behavior Therapy 1998;20:1–14.

Shaffer D, Fisher P, Lucas CP, Dulcan M, Schwab-Stone ME. NIMH Diagnostic Interview Schedule for Children, Version IV (NIMH DISC-IV): Description, differences from previous versions and reliability for some common diagnoses. Journal of the American Academy of Child and Adolescent Psychiatry 2000;39:28–38. [PubMed: 10638065]

Shapiro ES, Heick P. School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. Psychology in the Schools 2004;41:551–561.

Solanto MV, Gilbert SN, Raj A, Zhu J, Pope-Boyd S, Stepak B, Vail L, Newcorn JH. Neurocognitive functioning in AD/HD Predominantly Inattentive and Combined subtypes. Journal of Abnormal Child Psychology 2007;35:729–744. [PubMed: 17629724]

Teicher MH, Ito Y, Glod CA, Barber NI. Objective measurement of hyperactivity and attentional problems in ADHD. Journal of the American Academy of Child and Adolescent Psychiatry 35:334–342. [PubMed: 8714322]

Turkoglu, OD. Unpublished doctoral dissertation. University of Vermont; Burlington, VT: 2009. Reliability of test session observations and cross-informant agreement on children's behavioral and emotional problems.

Volpe RJ, DiPerna JC, Hintze JM, Shapiro ES. Observing students in classroom settings: A review of seven coding schemes. School Psychology Review 2005;34:454–474.

Volpe RJ, McConaughy SH, Hintze JM. Generalizability of classroom behavior problem and on-task scores for the Direct Observation Form. School Psychology Review 2009;39:382–401.

Wechsler, D. Wechsler Intelligence Scale for Children-Fourth Edition. San Antonio, TX: Psychological Corporation; 2003.

Wechsler, D. Wechsler Individual Achievement Tests-Second Edition. San Antonio, TX: Psychological Corporation; 2002.

Willcutt EG, Hartung CM, Lahey BB, Loney J, Pelham WE. Utility of behavior ratings by examiners during assessments of preschool children with Attention-Deficit/Hyperactivity Disorder. Journal of Abnormal Child Psychology 1999;27:463–472. [PubMed: 10821628]

Table 1
Sample Demographic Characteristics and DSM-IV-TR Diagnoses

Characteristics		**Subsamples for Logistic Regressions**		
	Primary Sample (*N* = 310)	**ADHD-C (*n* = 98)**	**ADHD-IN (*n* = 23)**	**Non-ADHD (*n* = 79)**
Boys, *n* (%)	215 (69.4)	63 (64.3)	18 (78.3)	51 (64.6)
Girls, *n* (%)	95 (30.6)	35 (35.7)	5 (21.7)	28 (35.4)
Mean Age (*SD*)	8.2 (1.6)	7.9 (1.6)	8.3 (1.5)	8.5 (1.6)
Mean SES (*SD*)[a]	5.9 (1.9)	5.5 (1.9)	6.4 (1.6)	6.7 (1.7)
Ethnicity, *n* (%)				
Non-Latino White	185 (59.7)	59 (60.2)	16 (69.6)	53 (67.1)
African American	91 (29.4)	32 (32.7)	4 (17.4)	14 (17.7)
Latino/Hispanic	17 (5.5)	2 (2.0)	2 (2.5)	4 (5.0)
Other or Unknown	17 (5.5)	5 (5.1)	1 (4.3)	8 (10.1)
DSM-IV-TR Diagnoses, *n* (%)[b]				
ADHD-Combined	106 (34.2)	81 (82.7)	0	0
ADHD-Inattentive	55 (17.7)	17 (17.3)	23 (100)	0
Conduct Disorder	25 (8.1)	15 (15.3)	1 (4.3)	0
Dysthymia or Major Depression	13 (4.2)	6 (6.1)	2 (8.7)	3 (3.8)
Generalized Anxiety Disorder	15 (4.8)	5 (5.1)	2 (8.7)	3 (3.8)
Obsessive Compulsive Disorder	12 (3.9)	3 (3.1)	1 (4.3)	4 (5.1)
Oppositional Defiant Disorder	109 (35.2)	56 (57.1)	7 (30.4)	17 (21.5)
Separation Anxiety	33 (10.6)	10 (10.2)	2 (8.7)	5 (6.3)
Social Phobia/Agoraphobia	14 (4.5)	3 (3.1)	3 (13.0)	4 (5.1)
Specific Phobia	71 (22.9)	24 (24.5)	6 (26.0)	14 (17.7)
Tourette's or Tic Disorder	21 (6.8)	4 (4.1)	1 (4.3)	4 (5.1)
No diagnosis, *n* (%)	90 (29.0)	0	0	44 (55.7)
One diagnosis, *n* (%)	90 (29.0)	32 (32.7)	11 (47.8)	23 (29.1)
Two or more diagnoses, *n* (%)	130 (41.9)	66 (67.7)	12 (52.7)	12 (15.2)

Note: ADHD= Attention Deficit/Hyperactivity Disorder; ADHD-C = ADHD-Combined type; ADHD-IN = ADHD-Inattentive type.

[a] Socioeconomic status (SES) scored on an adapted version of Hollingshead's (1975) scale where 1 = lowest and 9 = highest (*n* = 284).

[b] Children with comorbid diagnoses were counted more than once for the different diagnostic categories.

Table 2

Means, Standard Deviations, and Intercorrelations among CBCL, TRF, TOF, and DOF Empirically Based Syndromes and Dependent Measures

Predictors	1	2	3	4	5	6	7	8	9	10	11
1. CBCL Attention Problems	8.69 (4.90)										
2. TRF Attention Problems	.31***	24.34 (12.75)									
3. TOF Attention Problems-WISC-IV	.22***	.25***	8.70 (7.63)								
4. TOF Oppositional-WISC-IV	.21***	.23***	.68***	2.82 (4.56)							
5. DOF Attention Problems	.11	.20**	.15**	.09	5.54 (2.95)						
6. DOF Oppositional	.07	.32***	.19**	.18**	.47***	1.97 (2.30)					
7. DOF Intrusive	.07	.21***	.11	.10	.37***	.62***	2.11 (2.79)				
Dependent Measures											
8. ADHDRS-IV-Home Inattention	.81***	.33***	.20**	.19**	.10	.06	.04	13.77 (7.61)			
9. ADHDRS-IV-Home Hyperactivity-Impulsivity	.66***	.34***	.33***	.30***	.12*	.17**	.20***	.68***	11.60 (7.85)		
10. ADHDRS-IV-School Inattention	.37***	.83***	.20**	.19**	.13*	.23***	.15**	.39***	.30***	14.32 (7.97)	
11. ADHDRS-IV-School Hyperactivity-Impulsivity	.24***	.73***	.28***	.25***	.18**	.35***	.39***	.23***	.44***	.62***	11.21 (8.54)

Note: N = 310. Means and standard deviations (in parentheses) are on the diagonal. CBCL = Child Behavior Checklist; TRF = Teacher's Report Form; TOF = Test Observation Form; DOF = Direct Observation Form; ADHDRS-IV = Attention Deficit/Hyperactivity Disorder Rating Scale-IV.

[*] p <.05.

[**] p <.01.

[***] p <.001.

Table 3

Means, Standard Deviations, and Intercorrelations among CBCL, TRF, TOF, and DOF DSM-oriented Scales and Dependent Measures

Predictors	1	2	3	4
1. CBCL Attention Deficit/Hyperactivity Problems	7.09 (3.89)			
2. TRF Attention Deficit/Hyperactivity Problems-WISC-IV	.37***	13.94 (7.81)		
3. TOF Attention Deficit/Hyperactivity Problems	.30***	.30***	10.30 (9.47)	
4. DOF Attention Deficit/Hyperactivity Problems	.22***	.35***	.17**	8.20 (5.51)
Dependent Measures				
5. ADHDRS-IV-Home Inattention	.73***	.29***	.19**	.09
6. ADHDRS-IV-Home Hyperactivity-Impulsivity	.79***	.40***	.33***	.20***
7. ADHDRS-IV-School Inattention	.35***	.76***	.21***	.22***
8. ADHDRS-IV-School Hyperactivity-Impulsivity	.37***	.84***	.30***	.36***

Note: *N* = 310. Means and standard deviations (in parentheses) are on the diagonal. CBCL = Child Behavior Checklist; TRF = Teacher's Report Form; TOF = Test Observation Form; DOF = Direct Observation Form; ADHDRS-IV = Attention Deficit/Hyperactivity Disorder Rating Scale-IV. Correlations between ADHDRS-IV scales are shown in Table 1.

* *p* <.05.

** *p* <.01.

*** *p* <.001.

Table 4

Hierarchical Multiple Regressions of CBCL, TRF and TOF Scales Predicting Parent and Teacher Reports of Hyperactivity-Impulsivity

	Parent-rated Hyperactivity-Impulsivity[a]					Teacher-rated Hyperactivity-Impulsivity[b]				
Predictors	***B***	***R^2***	**ΔR^2**	***Adj.R^2***	***Semi-partial r^2***	***B***	***R^2***	**ΔR^2**	***Adj.R^2***	***Semi-partial r^2***
Model 1a										
Step 1		.46***		.45			.54***		.53	
CBCL Attention Problems	.62***				.34	.01				.00
TRF Attention Problems	.15**				.02	.73***				.48
Step 2		.49***	.03***	.48			.55***	.01**	.54	
CBCL Attention Problems	.59***				.30	.00				.00
TRF Attention Problems	.11^{*c}				.01	.71***				.44
TOF Attention Problems-WISC-IV	.18***				.03	.11**				.01
Model 1b										
Step 1		.46***		.45			.54***		.53	
CBCL Attention Problems	.62***				.34	.01				.00
TRF Attention Problems	.15**				.02	.73***				.48
Step 2		.48***	.02**	.47			.54***	.01	.54	
CBCL Attention Problems	.59***				.31	.00				.00
TRF Attention Problems	.12**				.01	.71***				.45
TOF Oppositional-WISC-IV	.15**				.02	.08^{e}				.01
Model 1c										
Step 1		.63***		.63			.71***		.71	
CBCL Attention Deficit/Hyperactivity Problems	.74***				.47	.06				.00
TRF Attention Deficit/Hyperactivity Problems	.13**				.01	.82***				.58
Step 2		.64***	.01*	.63			.72***	.00	.71	
CBCL Attention Deficit/Hyperactivity Problems	.72***				.43	.06				.00
TRF Attention Deficit/Hyperactivity Problems	.11**				.01	.81***				.54

	Parent-rated Hyperactivity-Impulsivity[a]					Teacher-rated Hyperactivity-Impulsivity[b]				
Predictors	***B***	R^2	ΔR^2	*Adj.* R^2	*Semi-partial* r^2	***B***	R^2	ΔR^2	*Adj.* R^2	*Semi-partial* r^2
TOF Attention Deficit/Hyperactivity Problems-WISC-IV	.08*[c, d]				.01	.04				.00

Note: N = 310. B = Standardized Beta weights; *Adj.* R^2 = Adjusted R^2; Semi-partial r^2 = unique variance accounted for by each predictor. CBCL = Child Behavior Checklist; TRF = Teacher's Report Form; TOF = Test Observation Form.

[a] From Attention Deficit/Hyperactivity Disorder Rating Scale-IV-Home version.

[b] From Attention Deficit/Hyperactivity Disorder Rating Scale-IV-School version

[c] Possible chance effect when corrected for number of analyses (Sakoda, Cohen & Beall, 1954).

[d] With removal of two outliers, Beta = .07, p = .065.

[e] With removal of one outlier, Beta = .08, p = .042.

[*] $p < .05$.

[**] $p < .01$.

[***] $p < .001$.

Table 5

Hierarchical Multiple Regressions of CBCL, TRF and DOF Scales Predicting Parent and Teacher Reports of Hyperactivity-Impulsivity

	Parent-rated Hyperactivity-Impulsivity[a]					Teacher-rated Hyperactivity-Impulsivity[b]				
Predictors	***B***	***R^2***	***ΔR^2***	***Adj.R^2***	***Semi-partial r^2***	***B***	***R^2***	***ΔR^2***	***Adj.R^2***	***Semi-partial r^2***
Model 2a										
Step 1		.46***		.45			.54		.53	
CBCL Attention Problems	.62***				.34	.01				.00
TRF Attention Problems	.15**				.02	.73***				.48
Step 2		.46***	.00	.45			.54	.00	.53	
CBCL Attention Problems	.62***				.34	.01				.00
TRF Attention Problems	.14**				.02	.72***				.46
DOF Attention Problems	.03				.00	.04				.00
Model 2b										
Step 1		.46***		.45			.54***		.53	
CBCL Attention Problems	.62***				.34	.01				.00
TRF Attention Problems	.15**				.02	.73***				.48
Step 2		.46**	.01	.46			.55***	.02**	.55	
CBCL Attention Problems	.62***				.34	.02				.00
TRF Attention Problems	.12^{*c}				.01	.69***				.38
DOF Oppositional	.08^{d}				.01	.13**				.02
Model 2c										
Step 1		.46***		.45			.54***		.53	
CBCL Attention Problems	.62***				.34	.01				.00
TRF Attention Problems	.15**				.02	.73***				.48
Step 2		.47***	.02**	.47			.60***	.06***	.59	
CBCL Attention Problems	.61***				.34	.01				.00
TRF Attention Problems	.12**				.01	.68***				.40

	Parent-rated Hyperactivity-Impulsivity^{a}					**Teacher-rated Hyperactivity-Impulsivity**^{b}				
Predictors	***B***	R^2	ΔR^2	***Adj.*** R^2	***Semi-partial*** r^2	***B***	R^2	ΔR^2	***Adj.*** R^2	***Semi-partial*** r^2
DOF Intrusive	.13**				.02	.25***				.06
Model 2d										
Step 1		.63***		.63			.71***		.71	
CBCL Attention Deficit/Hyperactivity Problems	.74***				.47	.06				.00
TRF Attention Deficit/Hyperactivity Problems	.13**				.01	.82***				.58
Step 2		.63***	.00	.63			.72***	.01*	.72	
CBCL Attention Deficit/Hyperactivity Problems	.74***				.47	.06				.00
TRF Attention Deficit/Hyperactivity Problems	.13**				.01	.80***				.50
DOF Attention Deficit/Hyperactivity Problems	.00				.00	.07$^{*\,c,\,e}$				.00

Note: N = 310. B = Standardized Beta weights; *Adj.* R^2 = Adjusted R^2; Semi-partial r^2 = unique variance accounted for by each predictor. CBCL = Child Behavior Checklist; TRF = Teacher's Report Form; DOF = Direct Observation Form.

[a] From Attention Deficit/Hyperactivity Disorder Rating Scale-IV-Home version.

[b] From Attention Deficit/Hyperactivity Disorder Rating Scale-IV-School version

[c] Possible chance effect when corrected for number of analyses (Sakoda, Cohen & Beall, 1954).

[d] With removal of two outliers, Beta = .09, p = .039.

[e] With removal of one outlier, Beta = .05, p = .104.

[*] $p < .05$.

[**] $p < .01$.

[***] $p < .001$.

Table 6

Multinomial Logistic Regressions of CBCL, TRF and TOF Scales for Predicting Categorical ADHD Diagnoses

Predictors	ADHD-C vs. Non-ADHD Odds Ratio[a]	ADHD-C vs. ADHD-IN Odds Ratio[b]	Nagelkerke R^2
Model 3a			.69***
1. CBCL Attention Problems	1.55 (1.32-1.83)***	1.03 (0.90-1.18)	
2. TRF Attention Problems	1.16 (1.10-1.23)***	1.04 (0.99-1.09)	
3. TOF Attention Problems-WISC-IV	1.12 (1.02-1.23)*[c]	1.08 (1.01-1.17)*	
Model 3b			.67***
1. CBCL Attention Problems	1.54 (1.31-1.81)***	1.03 (0.90-1.17)	
2. TRF Attention Problems	1.16 (1.11-1.23)***	1.04 (0.99-1.09)	
3. TOF Oppositional-WISC-IV	1.09 (0.95-1.25)	1.09 (0.96-1.24)	
Model 3c			.72***
1. CBCL Attention Deficit/Hyperactivity Problems	1.85 (1.49-2.30)***	1.28 (1.06-1.54)*[c]	
2. TRF Attention Deficit/Hyperactivity Problems	1.34 (1.21-1.48)***	1.15 (1.06-1.26)**	
3. TOF Attention Deficit/Hyperactivity Problems-WISC-IV	1.04 (0.97-1.12)	1.05 (0.98-1.21)	

Note: Total *N* = 200. ADHD-C = Attention Deficit/Hyperactivity Disorder-Combined type; ADHD-IN = Attention Deficit/Hyperactivity Disorder-Inattentive type; TOF = Test Observation Form; CBCL = Child Behavior Checklist; TRF = Teacher's Report Form.

[a] Odds ratios (95% confidence interval) showing independent contribution of each predictor with Non-ADHD as reference category.

[b] Odds ratios (95% confidence interval) showing independent contribution of each predictor with ADHD-IN as reference category.

[c] Possible chance effect when corrected for number of analyses (Sakoda, Cohen & Beall, 1954).

* p <.05.

** p <.01.

*** p <.001.

Table 7

Multinomial Logistic Regressions of CBCL, TRF, and DOF Scales for Predicting Categorical ADHD Diagnoses

	ADHD-C vs. Non-ADHD	ADHD-C vs. ADHD-IN	
Predictors	**Odds Ratio**[a]	**Odds Ratio**[b]	**Nagelkerke R^2**
Model 4a			.67**
1. CBCL Attention Problems	1.52 (1.30-1.78)***	1.02 (0.90-1.17)	
2. TRF Attention Problems	1.17 (1.10-1.24)***	1.05 (1.00-1.10)	
3. DOF Attention Problems	1.16 (.94-1.43)	1.00 (.84-1.19)	
Model 4b			.68***
1. CBCL Attention Problems	1.54 (1.32-1.81)***	1.03 (0.90-1.18)	
2. TRF Attention Problems	1.16 (1.10-1.23)***	1.03 (0.98-1.08)	
3. DOF Oppositional	1.17 (0.86-1.59)	1.58 (1.08-2.33)^{*c}	
Model 4c			.72***
1. CBCL Attention Problems	1.58 (1.33-1.86)***	1.03 (0.89-1.19)	
2. TRF Attention Problems	1.18 (1.11-1.25)***	1.04 (0.99-1.10)	
3. DOF Intrusive	1.83 (1.19-2.82)**	1.68 (1.14-2.47)**	
Model 4d			.73***
1. CBCL Attention Deficit/Hyperactivity Problems	1.86 (1.50-2.31)***	1.30 (1.08-1.57)**	
2. TRF Attention Deficit/Hyperactivity Problems	1.34 (1.20-1.48)***	1.16 (1.07-1.27)**	
3. DOF Attention Deficit/Hyperactivity Problems	1.17 (1.02-1.35)*	1.07 (0.96-1.21)	

Note: Total N = 200. ADHD-C = Attention Deficit/Hyperactivity Disorder-Combined type; ADHD-IN = Attention Deficit/Hyperactivity Disorder-Inattentive type; DOF = Direct Observation Form; CBCL = Child Behavior Checklist; TRF = Teacher's Report Form.

[a] Odds ratios (95% confidence interval) showing independent contribution of each predictor with Non-ADHD as reference category.

[b] Odds ratios (95% confidence interval) showing independent contribution of each predictor with ADHD-IN as reference category.

[c] Possible chance effect when corrected for number of analyses (Sakoda, Cohen & Beall, 1954).

* p <.05.

** p <.01.

***p <.001.

Colleagues to Receive Complimentary Copies of the *Journal of Clinical Child and Adolescent Psychology*

George Bear, Ph.D. 221A Willard Hall Education Building University of Delaware Newark, DE 19716	William R. Jenson, Ph.D. Department of Educational Psychology University of Utah 327 Milton Bennison Hall Salt Lake City, UT 84112
Sandra M. Chafouleas, Ph.D. Dept. of Educational Psychology Neag School of Education 249 Glenbrook R. (U-2064) Storrs, CT 06269	Kenneth W. Merrell, Ph.D. Department Head, School Psychology School Psychology 5208 University of Oregon Eugene, OR 97403-5261
George DuPaul, Ph.D. Lehigh University Iacocca Hall 111 Research Drive Bethlehem, PA 18015	Thomas J. Power, Ph.D. Center for Management of ADHD Children's Hospital of Philadelphia 3405 Civic Center Blvd. Philadelphia, PA 19104-4399
Randy Floyd, Ph.D. Department of Psychology University of Memphis Memphis, TN 38152	Timothy Stickle, Ph.D. Psychology Department University of Vermont John Dewey Hall Burlington, VT 05401
Betsy Hoza, Ph.D. Psychology Department University of Vermont John Dewey Hall Burlington, VT 05401	Robert J. Volpe, Ph.D. Department of Counseling & Applied Educational Psychology 203A Lake Hall Northeastern University Boston, MA 02115-5000