

Standardized Observational Assessment of Attention Deficit Hyperactivity Disorder Combined and Predominantly Inattentive Subtypes. II. Classroom Observations

Stephanie H. McConaughy and Masha Y. Ivanova
University of Vermont

Kevin Antshel
SUNY Upstate Medical University

Ricardo B. Eiraldi
University of Pennsylvania

Levent Dumenci
Virginia Commonwealth University

Abstract. Trained classroom observers used the Direct Observation Form (DOF; McConaughy & Achenbach, 2009) to rate observations of 163 6- to 11-year-old children in their school classrooms. Participants were assigned to four groups based on a parent diagnostic interview and parent and teacher rating scales: Attention Deficit Hyperactivity Disorder (ADHD)—Combined type ($n = 64$); ADHD—Inattentive type ($n = 22$); clinically referred without ADHD ($n = 51$); and nonreferred control children ($n = 26$). The ADHD—Combined group scored significantly higher than the referred without ADHD group and controls on the DOF Intrusive and Oppositional syndromes, Attention Deficit Hyperactivity Problems scale, Hyperactivity-Impulsivity subscale, and Total Problems; and significantly lower on the DOF On-Task score. The ADHD—Inattentive group scored significantly higher than controls on the DOF Sluggish Cognitive Tempo and Attention Problems syndromes, Inattention subscale, and Total Problems; and significantly lower on the DOF On-Task score. Implications are discussed regarding the discriminative validity of standardized classroom observations for identifying children with ADHD and differentiating between the two ADHD subtypes.

According to the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; DSM-IV; American Psychiatric Association, 1994) and its text revision (DSM-IV-TR;

Preparation of this article was supported in part by Grant R01 HD40220 from the National Institute of Child Health and Human Development to Stephanie H. 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. The first author has a potential conflict of interest because she is a developer of the Direct Observation Form, the primary measure used in this study. However, the second author managed data collection and data analyses

American Psychiatric Association, 2000), approximately 3%–7% of the school-age children in the United States meet diagnostic criteria for attention deficit hyperactivity disorder (ADHD). The DSM-IV-TR lists 18 behavioral symptoms of ADHD, with 9 symptoms clustered under Inattention and 9 clustered under Hyperactivity-Impulsivity. These groupings of symptoms are then used to define three subtypes of the disorder: Predominantly Inattentive (ADHD-IN), showing 6 of 9 symptoms of inattention, but fewer than 6 symptoms of hyperactivity and impulsivity; Predominantly Hyperactive-Impulsive, showing the opposite symptom pattern; and Combined (ADHD-C), showing at least 6 of 9 symptoms of both inattention and hyperactivity-impulsivity. To meet diagnostic criteria for any subtype, the DSM-IV-TR requires persistence of symptoms for at least 6 months; age of onset before age 7 for at least some of the symptoms; functional impairment from the symptoms in two or more settings (e.g., home, school, or work); and evidence of clinically significant functional impairment in social, academic, or occupational functioning.

Prevalence rates for the ADHD subtypes, as well as ADHD in general, have varied depending on the research criteria used for assessing symptoms. Across five international studies that used structured parent and/or teacher ratings plus impairment criteria, Nigg (2006) computed prevalence rates of 2.9% for ADHD-C, 3.2% for ADHD-IN, and 0.6% for Predominantly Hyperactive-Impulsive ADHD, and a total of 6.8% for any ADHD. Applying prevalence rates of 3%–7% to school settings, at least 1–2 children in an elementary classroom of 20–30 students might be expected to exhibit symptoms of ADHD.

Numerous studies have shown that parent and teacher ratings of the core symptoms (inattention, hyperactivity, and impulsivity) differentiate children with ADHD from those without ADHD (for reviews, see Barkley, 2006; DuPaul & Stoner, 2003; Nigg, 2006). Research has also shown that the level of agreement between parents and teachers regarding children's problems is moderate at best, with correlations ranging from .21 to .44,

depending on the nature of the problems (Achenbach, McConaughy, & Howell, 1987; Achenbach & Rescorla, 2001; Mitsis, McKay, Schulz, Newcorn, & Halperin, 2000). In addition, efforts to develop laboratory tests or neuropsychological measures of ADHD have produced limited results, particularly for identifying the ADHD subtypes (e.g., Solanto et al., 2007).

Considering the moderate agreement between parent and teacher reports and the absence of a definitive test for ADHD, direct observations of children's classroom behavior may be one venue for external validation of ADHD symptoms. In a review of 39 observational studies, Platzman, Stoy, Brown, Coles, Smith, and Falek (1992) found, in general, that classroom observations were better than laboratory observations for distinguishing children with ADHD from comparison groups without ADHD. Behaviors that most consistently identified children with ADHD were measures of attention (including off-task behavior), activity level, vocalization, and for some studies, negative interpersonal interactions.

Direct observations of children's classroom behavior are among the most common assessment methods used by school psychologists (Shapiro & Heick, 2004; Wilson & Reschly, 1996). Moreover, 94% of surveyed school psychologists said they conducted classroom observations as one of their methods for assessing ADHD (Demaray, Schaefer, & Delong, 2003). Volpe, DiPerna, Hintze, and Shapiro (2005) reviewed seven systematic coding systems that school psychologists

with no conflict of interest. We are grateful to Thomas Achenbach and Robert Volpe for their contributions to this work.

Correspondence regarding this article should be addressed to Stephanie H. McConaughy, Department of Psychiatry, University of Vermont, One South Prospect Street, Burlington, VT 05401; E-mail: stephanie.mcconaughey@uvm.edu

Copyright 2009 by the National Association of School Psychologists, ISSN 0279-6015, which has nonexclusive ownership in accordance with Division G, Title II, Section 518 of P.L. Law 110-161 and NIH Public Access Policy

might use to observe children's behavior in school classrooms. Three observation systems, in particular, showed good reliability and validity for ADHD: the Attention-Deficit Hyperactivity Disorder School Observation Code (Gadow, Sprafkin, & Nolan, 1996); Behavior Observation of Students in Schools (Shapiro, 2004); and Classroom Observation Code (Abikoff & Gittelman, 1985). These three observation systems all relied on observations of the occurrence or nonoccurrence of discrete behaviors over short intervals of time. Although they provide reliable and valid indices of observable ADHD-consistent behaviors, all three coding systems have the disadvantage of focusing only on a limited number of observable behaviors at any one time.

Taking a different approach, we combined the advantages of systematic direct observations and behavior rating scales to assess an array of observable behaviors in school classrooms. To do this, we used a newly revised version of the Direct Observation Form (DOF; McConaughy & Achenbach, 2009), which is a standardized rating form developed as part of the Achenbach System of Empirically Based Assessment (Achenbach & Rescorla, 2001). The DOF can be used to obtain multiple 10-min observations of a child's classroom behavior. During each 10-min observation, the observer records a narrative description of observed behaviors and codes occurrences and nonoccurrences of on-task behavior in ten 1-min intervals. Immediately following each 10-min observation, the observer rates the child on 89 problem items, using a 4-point scale. Item ratings are then averaged across multiple observation sessions to obtain summative scale scores for five empirically based syndrome scales (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. On-task occurrences are summed for each 10-min observation and then averaged across multiple observations (for details, see Method section).

Several studies have demonstrated good reliability and validity of an earlier version of the DOF (Achenbach, 1986; McConaughy, Achenbach, & Gent, 1988; McConaughy, Kay, & Fitzgerald, 1999; Reed & Edelbrock, 1983). We know of only one previous study that used the DOF to compare observations of children with ADHD symptoms versus controls in the same classrooms. In that study, Skansgaard and Burns (1998) added 9 new items to the 97 items of the 1986 DOF to create problem scales for inattention, hyperactivity-impulsivity, oppositional defiant disorder/overt conduct disorder (ODD/overt CD), and what they termed "slow cognitive tempo." Inter-rater reliabilities ranged from .69 to .97 for the four scales, plus 1.00 for the DOF On-Task score. To test the discriminative validity of the DOF, Skansgaard and Burns grouped their sample into an ADHD-C subtype ($n = 6$), ADHD-IN subtype ($n = 6$), and matched controls ($n = 12$), based on percentile cut points for teachers' ratings of DSM-IV ADHD symptoms. Despite the small sample sizes, the ADHD-C and ADHD-IN groups both scored significantly higher than controls on the DOF Inattention scale and lower on DOF On-Task, and the ADHD-IN group scored significantly higher than controls on the DOF Slow Cognitive Tempo scale. The ADHD-C group scored significantly higher than the ADHD-IN group and controls on the DOF Hyperactivity-Impulsivity and ODD/overt CD scales. Similar group differences were reported for teachers' ratings on comparable subsets of DSM-IV symptoms, except that ADHD-IN scored significantly higher than both ADHD-C and controls on Slow Cognitive Tempo.

Purpose of the Present Study

In the present study, independent observers used the 2009 DOF to rate classroom behaviors for children with DSM-IV-TR diagnoses of ADHD-C and ADHD-IN, plus clinically referred children without ADHD (NON-ADHD REF) and typically developing nonreferred controls (Control). Observers were blind to children's group assignment and as-

assessment results for the study. We had the following hypotheses: (a) Children with ADHD-C would score significantly higher than NON-ADHD REF and Control on the DOF Attention Problems syndrome, the DSM-oriented Attention Deficit Hyperactivity Problems scale, and the Inattention and Hyperactivity-Impulsivity subscales, and lower on DOF On-Task. (b) Children with ADHD-IN would score significantly higher than NON-ADHD REF and Control on the DOF Attention Problems and Sluggish Cognitive Tempo syndromes, Attention Deficit Hyperactivity Problems and the Inattention subscale, and lower on DOF On-Task. (c) Children with ADHD-C would score significantly higher than children with ADHD-IN on Attention Problems and Attention Deficit Hyperactivity Problems (both of which include hyperactivity and impulsivity along with inattention), and the Hyperactivity-Impulsivity subscale. Based on findings from Skansgaard and Burns (1998), we also hypothesized that children with ADHD-C would score significantly higher than the other three groups on the DOF Oppositional syndrome. Because children with both ADHD subtypes should have problems with inattention, we expected no significant differences between ADHD-C and ADHD-IN on the Inattention subscale or On-Task score. We made no a priori hypotheses regarding differences between children with ADHD and NON-ADHD REF and Control on other DOF problem scales. We limited our research to ages 6–11 to tap ADHD symptoms prior to adolescence.

Method

Participants

This study was part of a larger federally funded research effort to test the potential contributions of standardized observations to assessment of ADHD. Participants were recruited from mental health providers and public and private schools in the vicinity of 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; $n =$

53); The Children's Hospital of Philadelphia in Pennsylvania (CHOP, $n = 46$); and the Child and Adolescent Psychiatry Clinic at SUNY Upstate Medical University in Syracuse, New York (SUNY, $n = 64$). The UVM clinic was in a small urban area and the CHOP and SUNY clinics were in large urban centers. Participants were drawn from a total sample of 456, 6- to-11-year-old children participating in the larger study. The participants were also a subset of the sample used by McConaughy, Ivanova, Antshel, and Eiraldi (2009) to test group differences in test session observations.

Diagnostic Group Assignment

To be assigned to the ADHD-C group, the child had to have symptoms of inattention and hyperactivity-impulsivity based on combined parent and teacher reports. Specifically, the child had to have a positive diagnosis of ADHD-Combined type (314.01) on the Diagnostic Report of the National Institute of Mental Health Diagnostic Interview Schedule for Children—Fourth Edition (NIMH DISC-4; Shaffer et al., 2000), plus scores ≥ 80 th percentile on the Inattention or Hyperactivity-Impulsivity subscales of the ADHD Rating Scale—IV: School version (ADHDRS-IV-School version; DuPaul, Power, Anastopolous, & Reid, 1998); or the child had to have a positive diagnosis of ADHD-Predominantly Inattentive type (314.00) on the NIMH DISC-4 Diagnostic Report, plus scores ≥ 80 th percentile on both the Inattention and Hyperactivity-Impulsivity subscales of the ADHDRS-IV-School version (see Measures section for descriptions of instruments). To be assigned to the ADHD-IN group, the child had to have a DSM-IV-TR diagnosis of ADHD-Predominantly Inattentive type (314.00) on the NIMH DISC-4 Diagnostic Report, plus a score ≥ 80 th percentile on the Inattention subscale of the ADHDRS-IV-School version and a score < 80 th percentile on the Hyperactivity-Impulsivity subscale of the ADHDRS-IV-School version. To be assigned to the NON-ADHD-REF group, the child had to have been referred to the study because of behavioral/emotional or learning problems, have no diagnosis of ADHD on the NIMH DISC-4, and have scores < 80 th percentile on the Inatten-

Table 1
Demographic Characteristics and DSM-IV-TR Diagnoses for Four Groups

Characteristic	ADHD-C (<i>n</i> = 64)	ADHD-IN (<i>n</i> = 22)	NON-ADHD-REF (<i>n</i> = 51)	Control (<i>n</i> = 26)
Boys, <i>n</i> (%)	49 (76.6%)	17 (77.3%)	35 (68.6%)	15 (57.7%)
Girls, <i>n</i> (%)	15 (23.4%)	5 (22.7%)	16 (31.4%)	11 (42.3%)
Mean age (<i>SD</i>)	7.5 (1.5)	8.3 (1.5)	8.5 (1.7)	8.6 (1.5)
Mean SES (<i>SD</i>) ^a	5.7 (1.8)	6.2 (1.8)	6.9 (1.6)	6.5 (1.8)
Ethnicity, <i>n</i> (%)				
Non-Latino White	38 (59.4%)	15 (68.2%)	38 (74.5%)	14 (53.8%)
African American	21 (32.8%)	4 (18.2%)	5 (9.8%)	8 (30.8%)
Latino/Hispanic	2 (3.1%)	2 (9.1%)	2 (3.9%)	1 (3.8%)
Other or Unknown	3 (4.7%)	1 (4.5%)	6 (11.9%)	3 (11.5%)
DSM-IV-TR Diagnoses, <i>n</i> (%) ^b				
ADHD only	20 (31.3%)	12 (54.5%)	0 (0%)	0 (0%)
Conduct disorder	11 (17.2%)	1 (4.5%)	0 (0%)	0 (0%)
Dysthymia or major depression	5 (7.8%)	2 (9.1%)	3 (5.9%)	0 (0%)
Generalized anxiety disorder	4 (6.3%)	2 (9.1%)	3 (5.9%)	0 (0%)
Obsessive compulsive disorder	2 (3.1%)	1 (4.5%)	3 (5.9%)	1 (3.8%)
Oppositional defiant disorder	33 (51.6%)	6 (27.3%)	14 (27.5%)	1 (3.8%)
Separation anxiety	12 (18.8%)	2 (9.1%)	5 (9.8%)	0 (0%)
Social phobia	2 (3.1%)	2 (9.1%)	3 (5.9%)	0 (0%)
Specific phobia	19 (29.7%)	5 (22.7%)	9 (17.6%)	6 (23.1%)
Tourette's or tic disorder	3 (4.7%)	1 (4.5%)	4 (7.8%)	0 (0%)
No diagnosis, <i>n</i> (%)	0 (0%)	0 (0%)	25 (49.0%)	19 (73.1%)
One diagnosis, <i>n</i> (%)	20 (31.3%)	12 (54.5%)	15 (29.4%)	6 (23.1%)
Two or more diagnoses, <i>n</i> (%)	44 (68.8%)	10 (45.5%)	11 (21.6%)	1 (3.8%)

Note. Total *N* = 163. DSM-IV-TR = *Diagnostic and Statistical Manual* (4th ed., text rev.); SES = socioeconomic status; ADHD-C = attention deficit hyperactivity disorder—Combined type; ADHD-IN = ADHD—Inattentive type; NON-ADHD-REF = non-ADHD clinically referred; Control = nonreferred controls.

^aSES scored on an adapted version of Hollingshead's (1975) scale for parental occupation where 1 = *lowest* and 9 = *highest*.

^bChildren with comorbid diagnoses were counted more than once for the different diagnostic categories.

tion and Hyperactivity-Impulsivity subscales of the ADHDRS-IV-School version. To be assigned to the Control group, the child had to have no diagnosis of ADHD on the NIMH DISC-4 and scores <80th percentile on the Inattention and Hyperactivity-Impulsivity subscales of the ADHDRS-IV-School version and the ADHDRS-IV: Home version (ADHDRS-IV-Home version). Additional criteria for recruitment of Control children were: did not repeat a grade and had not been referred for or received special education, a Section 504 plan, counseling, or mental health services within the past 12 months.

Exclusionary criteria for children in all four groups were: Wechsler Intelligence Scale for Children—Fourth Edition Full Scale IQ < 80, parent-reported physical or medical problems (e.g., seizure disorders, cerebral palsy), mental retardation, autism, or pervasive developmental disorder. No children in this study were taking medications for ADHD. Table 1 shows demographic characteristics of the sample.

As shown in Table 1, there were approximately three times as many boys than girls in the ADHD-C and ADHD-IN groups, consistent with rates reported in the DSM-IV-TR

(American Psychological Association, 2000). A one-way analysis of variance (ANOVA) showed significant group differences in age, $F(3,162) = 5.38, p = .001$, with ADHD-C being significantly younger than NON-ADHD REF and Control ($p < .05$). There were no significant site differences in age. Socioeconomic status (SES) was scored on an adaptation of Hollingshead's (1975) scale for occupation of the parent obtaining the higher score. The mean SES for the sample for whom SES was known was 6.3 ($SD = 1.8; n = 153$). A one-way ANOVA showed significant group differences in SES, $F(3,152) = 3.93, p = .01$, with ADHD-C scoring significantly lower than NON-ADHD REF ($p < .05$). There were also site differences in SES, $F(2,152) = 12.47, p < .001$, with CHOP cases having significantly lower SES (mean = 5.3, $SD = 1.7$) than UVM cases (mean = 6.9, $SD = 1.7$) and SUNY cases (mean = 6.5, $SD = 1.6; p < .05$). CHOP also had higher percentages of African American participants. Ethnicity of the total sample was 64.4% non-Latino White, 23.3% African American, 4.3% Latino or Hispanic, and 7.9% other or unknown.

The bottom half of Table 1 shows the number of participants with various DSM-IV-TR diagnoses, according to Diagnostic Reports obtained from the NIMH DISC-4. Children in the two ADHD groups were allowed to have other comorbid DSM-IV-TR diagnoses. Children in the NON-ADHD REF and Control groups were allowed to have diagnoses other than ADHD.

Measures

ADHDRS-IV. The ADHDRS-IV (DuPaul et al., 1998) is an 18-item rating scale, with 9 terms that assess DSM-IV defined symptoms of inattention and 9 terms that assess DSM-IV defined symptoms of hyperactivity-impulsivity. Each item is rated on a 4-point scale: 0 = *not at all, rarely*; 1 = *sometimes*; 2 = *often*; and 3 = *very often*. 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 ranging from .86 to .96 and test-retest reliabilities ranging from .78 to .90 over a 4-week interval. For scores ≥ 80 th percentile on the Hyperactivity-Impulsivity subscale of the ADHDRS-IV-School version, Power et al. (1998) reported a positive predictive probability of .75 and negative predictive probability of .87 for predicting ADHD-C versus Controls, when teacher ratings were combined with parent ratings. For scores ≥ 80 th percentile on the Inattention subscale, Power et al. reported a positive predictive probability of .65 and negative predictive probability of .90 for predicting ADHD-IN versus controls.

NIMH DISC-4. The NIMH DISC-4 (Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000) is a highly structured diagnostic interview administered to parents to assess criteria for DSM-IV-TR disorders applicable to children ages 6–17. Diagnoses are assessed for the past 12 months and past 4 weeks. For this study, we used the diagnostic reports obtained from the computer-assisted NIMH DISC-4 modules for ADHD, CD, ODD, anxiety disorders, and mood disorders. Shaffer et al. (2000) reported NIMH DISC-4 test-retest kappas of .96 for specific phobia, .79 for ADHD, .66 for major depression, .65 for generalized anxiety, .58 for separation anxiety, .54 for ODD and social phobia, and .43 for CD.

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-min 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 s of each 1-min interval (a predominant time-sampling method). Immediately after each 10-min observation, the observer rates the child on 89 problem items, using a 4-point scale: 0 = *no occurrence*; 1 = *very slight or ambig-*

uous 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. Examples of DOF items are as follows: (3) "Argues"; (7) "Doesn't concentrate or doesn't pay attention for long"; (9) "Doesn't sit still, restless, or hyperactive"; and (70) "Underactive, slow moving, or lacks energy." Item 89 is open-ended for rating other problems not covered by Items 1–88.

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), plus a DSM-oriented Attention Deficit Hyperactivity Problems scale and Inattention and Hyperactivity-Impulsivity subscales and Total Problems. The DOF syndrome scales were derived from exploratory and confirmatory factor analyses of classroom observations of 1,261 6- to 12-year-old children (McConaughy & Achenbach, 2009). Items with factor loadings ≥ 0.20 and $p < .01$ were retained on the syndrome scales. The root mean square error of approximation for the final five-factor solution was 0.024, which was well below the upper limit of 0.05 to 0.07 considered to indicate good fit (Brown & Cudeck, 1993; Yu & Muthén, 2002). The Attention Deficit Hyperactivity Problems scale includes 23 problem items consistent with a DSM-IV-TR diagnosis of ADHD. These 23 items are divided into 10 items for the Inattention subscale and 13 items for the Hyperactivity-Impulsivity subscale.

For the DOF problem scales, the 0–1–2–3 item ratings are averaged across multiple 10-min observation sessions and then summed to obtain a total raw score for each scale. The DOF On-Task score is the total number of 1-min intervals when the child was rated as on task, averaged across multiple 10-min observations. The DOF On-Task score thus ranges from 0 to 10. The DOF computer-scored profile provides raw scores for each scale plus *T* scores and percentiles based on separate norms for boys and girls ages 6–11 (for items comprising each DOF scale, see McConaughy

and Achenbach, 2009, and Volpe, McConaughy, and Hintze, 2009).

McConaughy and Achenbach (2009) reported internal consistency alphas from .49 to .87 for the nine DOF problem scales, with mean $\alpha = .74$. Inter-rater reliabilities for the nine problem scales and On-Task ranged from .71 to .97, averaged across 12 pairs of observers, as described later in the section on DOF training procedures.¹ Volpe et al. (2009) also reported generalizability and dependability coefficients $\geq .80$ for the DOF Intrusive, Oppositional, Attention Deficit Hyperactivity Problems, Hyperactivity-Impulsivity, and Total Problems scales. Criterion-related validity of the DOF was demonstrated by significantly ($p < .05$) higher scores for clinically referred than nonreferred 6- to 11-year-old children on all DOF scales (McConaughy & Achenbach, 2009). For the present study, we used a research edition of the DOF that included 115 problem items. However, data analyses were conducted using only the 89 items of the final 2009 version of the DOF.

Procedure

Recruitment of participants. The research protocol was approved by the Institutional Review Boards of each of the three sites. McConaughy et al. (2009) provide details on recruitment of participants, which are summarized here. To recruit participants for the three clinically referred groups (ADHD-C, ADHD-IN, NON-ADHD REF), researchers gave mental health clinicians and school personnel packets of letters and consent forms describing the goals and procedures of the study to parents. In order not to bias 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." Parents mailed consent forms directly to the research staff, who then scheduled appointments for testing the child at the clinic. Researchers contacted school staff to arrange observations of the child in the classroom. Parents of referred children were paid \$15 for their participation.

To recruit Control children, researchers gave school personnel packets of study materials containing written instructions and exclusionary criteria for selecting typically developing children. School staff randomly selected 3 boys and 2 girls per classroom, and then mailed letters and consent forms to their parents. Parents returned signed consent forms directly to the research staff. Parents of Control children were paid \$50 because they had less to gain from the psychological assessment battery than did parents of referred children.

So as not to create potential bias toward particular diagnoses, teachers were informed that the child was “participating in a study of children’s behavioral development.” Teachers were kept blind to the diagnostic group assignment of participants throughout the study. Teachers were paid \$15 for completing rating forms and allowing observations in their classrooms.

DOF training procedures. There were 24 observers across the three study sites, including undergraduate psychology students and postgraduates with bachelor’s or master’s degrees. Prior to observations for the present study, researchers trained observers in the DOF recording and rating procedures, as now described in the DOF manual (McConaughy & Achenbach, 2009). Following initial training in the rating rules for the problem items, each observer was paired with another observer to rate 5 anonymous children in local elementary schools as practice cases. The two observers then met with the trainer to discuss their ratings of each child and resolve any discrepancies. After completing the practice cases, the same observer pairs simultaneously observed 14–24 additional anonymous children. The number of 10-min observations per child varied across observer pairs. Nine observer pairs completed 1 DOF per child per observer, while three observer pairs completed 2–4 DOFs per child per observer. Observers were instructed not to discuss their ratings with each other until after data collection was completed on all observed children. Trainers then met again with each observer pair to discuss their ratings and clarify rating rules.

To obtain inter-rater reliabilities, we computed Pearson r values between raw scale scores separately for the nine DOF problem scales and On-Task for each of the 12 trained observer pairs for a total sample of 212 6- to 11-year-old children. We converted each r to Fisher’s z and then obtained a mean z and mean r for each DOF scale across the 12 observer pairs. Inter-rater reliabilities were .71 to .87 for the five syndrome scales, .80 for Attention Deficit Hyperactivity Problems, .70 for the Inattention subscale, .81 for the Hyperactivity-Impulsivity subscale, .88 for Total Problems, and .97 for On-Task.

Observation procedures. After training, observers used the DOF to obtain four 10-min observations of each child participant. (Ten of the 163 participants had three 10-min observations.) Observations were conducted on 2 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; 75th quartile = 4). 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-min observation period.

Observers followed the procedures described in the Measures section for recording observations of the child’s behavior. For ratings of on task, observers were instructed to consider a child to be on task if that child was doing what was expected in that situation for the majority of the last 5 s of each 1-min interval. Observers were given written exemplars of on-task behaviors (e.g., listening to the teacher’s directions; reading a book; working on an assigned task at desk; listening to others in circle time) and off-task behaviors (e.g., doing something that requires the teacher to redirect him or her to get back on task; doodling, drawing, or playing with a toy when supposed to be working; looking around the room or not looking at the teacher or someone else who is speaking).

Immediately following each 10-min observation, observers rated each child on the 89

DOF problem items, using the 4-point scale described in the Measures section. Observers were provided written rules and exemplars for choosing among the DOF problem items and deciding whether to rate items 0, 1, 2, or 3, as described in the DOF manual (McConaughy & Achenbach, 2009). Observers were blind to all information about the child, including referral complaints, test scores, parent and teacher rating scale scores, and the child's diagnostic group assignment. Observers were also instructed not to discuss their observations with teachers.

Data Analyses

To test mean differences on the DOF scales, we performed two separate 2×4 multivariate ANOVAs (MANOVA), treating gender and diagnostic group (ADHD-C, ADHD-IN, NON-ADHD REF, and Control) as between-subject variables and summative raw scores on the five DOF syndrome scales or Inattention and Hyperactivity-Impulsivity subscales as dependent variables (SPSS 15.0 general linear model). Each MANOVA was followed by a univariate ANOVA and post hoc Tukey honestly significant difference (Tukey HSD) tests to examine group differences. The Tukey HSD tests were performed on homogeneous subsets to adjust for unequal sample sizes.² We also performed two 2×4 ANOVAs on summative raw scores for DOF Total Problems, Attention Deficit Hyperactivity Problems, and On-Task, followed by Tukey HSD tests. We examined effect sizes (ES) indicated by partial η^2 , which can be translated directly into percentage of variance accounted for. According to Cohen's (1988) criteria: ES accounting for 1%–5.8% of variance are small; 5.9%–13.7% of variance are medium; and >13.8% of variance are large.

We performed discriminant analyses to determine which combinations of the DOF scales contributed to discriminating between the following groups: (a) ADHD-C versus NON-ADHD REF, (b) ADHD-IN versus NON-ADHD REF, (c) ADHD-IN versus Control, (d) ADHD-C versus Control, and (e) ADHD-C versus ADHD-IN. For each discriminant analysis,

we treated the following variables as sets of candidate predictors: (a) five DOF syndromes, (b) five DOF syndromes plus On-Task, (c) Attention Deficit Hyperactivity Problems plus On-Task, (d) the Inattention and Hyperactivity-Impulsivity subscales plus On-Task, and (e) Total Problems plus On-Task. We entered all candidate predictors simultaneously. For completeness, we obtained cross-validated classification rates in two ways: (a) setting prior probabilities as equal for both groups and (b) computing prior probabilities from group sizes. The magnitude of the standardized canonical coefficients can be interpreted as indicating the relative importance of each discriminating variable to predicting group membership (regardless of sign).

Results

Group Differences on DOF Scales

DOF. Table 2 shows means and standard deviations for the DOF scales for the four diagnostic groups. The overall MANOVA for the five DOF syndromes showed a significant main effect of group, $F(15, 417) = 4.04, p < .001$, partial $\eta^2 = .12$. Subsequent ANOVAs revealed significant group effects ($p < .01$) for all five syndrome scales, with medium to large ES (partial $\eta^2 = .08$ to $.14$). The 2×4 MANOVA on the five DOF syndromes also showed a significant gender effect, $F(5, 151) = 2.63, p = .026$, partial $\eta^2 = .08$; and a significant Gender \times Group interaction, $F(15, 417) = 2.00, p = .014$, partial $\eta^2 = .06$. Subsequent ANOVAs revealed a significant gender effect for the Immature/Withdrawn syndrome, $F(1, 155) = 4.16, p = .043$, partial $\eta^2 = .03$; and significant Group \times Gender interactions for the Sluggish Cognitive Tempo syndrome, $F(3, 155) = 4.67, p = .004$, partial $\eta^2 = .08$; and Immature/Withdrawn syndrome, $F(3, 155) = 5.66, p = .001$, partial $\eta^2 = .10$. On both syndromes, girls in the ADHD-IN group scored significantly higher ($p < .05$) than boys, in contrast to no differences between boys and girls in the other three groups.

The 2×4 ANOVA for the Attention Deficit Hyperactivity Problems scale revealed a significant main effect of group, with a large

Table 2
Group Differences on DOF Scales

DOF Scales	ADHD-C (<i>n</i> = 64)	ADHD-IN (<i>n</i> = 22)	NON-ADHD-REF (<i>n</i> = 51)	Control (<i>n</i> = 26)	<i>F</i>	<i>p</i>	Partial η^2
Empirically based syndromes							
Sluggish Cognitive Tempo	1.28 (1.26) _{a,b}	1.90 (1.72) _a	1.24 (1.35) _{a,b}	0.78 (0.86) _b	<i>F</i> (3, 155) = 6.63	<.001	.11
Immature/Withdrawn	0.88 (1.06) _a	0.75 (1.07) _{a,b}	0.45 (0.55) _{a,b}	0.27 (0.54) _b	<i>F</i> (3, 155) = 8.15	<.001	.14
Attention Problems	5.84 (2.65) _{a,b}	6.29 (2.82) _a	4.63 (2.45) _{a,b}	4.20 (2.61) _b	<i>F</i> (3, 155) = 4.42	.005	.08
Intrusive	3.25 (3.64) _a	1.00 (0.98) _b	0.84 (1.16) _b	1.08 (1.20) _b	<i>F</i> (3, 155) = 7.53	<.001	.13
Oppositional	2.94 (3.37) _a	1.30 (1.07) _b	1.47 (1.30) _b	0.75 (0.87) _b	<i>F</i> (3, 155) = 4.81	.003	.09
DSM-oriented scales							
Attention Deficit Hyperactivity Problems	10.40 (6.70) _a	7.87 (3.93) _{a,b}	5.88 (3.09) _b	4.92 (3.86) _b	<i>F</i> (3, 155) = 8.59	<.001	.14
Inattention subscale	3.51 (2.98) _a	3.17 (2.19) _a	2.34 (1.93) _{a,b}	1.52 (1.68) _b	<i>F</i> (3, 155) = 5.14	.002	.09
Hyperactivity-Impulsivity subscale	6.89 (4.45) _a	4.70 (2.51) _b	3.55 (1.91) _b	3.39 (2.48) _b	<i>F</i> (3, 155) = 9.02	<.001	.15
Total Problems	16.50 (9.51) _a	12.88 (5.83) _{a,b}	10.00 (4.62) _{b,c}	8.25 (5.79) _c	<i>F</i> (3, 155) = 9.99	<.001	.16
On-Task	6.76 (2.19) _a	7.76 (1.04) _{a,b}	8.24 (1.60) _{b,c}	8.89 (1.39) _c	<i>F</i> (3, 155) = 9.13	<.001	.15

Note. ADHD-C = Attention deficit hyperactivity disorder—Combined type; ADHD-IN = ADHD-Inattentive type; NON-ADHD-REF = non-ADHD clinically referred; Control = typically developing controls; DOF = Direct Observation Form; DSM = *Diagnostic and Statistical Manual* (4th ed., text rev.); HSD = honestly significant difference. Total *N* = 163. Mean (*SD*). DOF scale scores are raw scores averaged across 3 to 4 DOFs per case. On-task scores can range from 0 to 10. Means in the same row with the same subscripts do not differ significantly at *p* < .05 by Tukey HSD comparisons of homogeneous subsets.

ES (partial $\eta^2 = .14$). The 2×4 MANOVA for the Inattention and Hyperactivity-Impulsivity subscales showed a significant main effect of group, $F(6, 308) = 5.55$; $p < .001$; partial $\eta^2 = .10$, and subsequent ANOVAs revealed significant group effects ($p < .01$) for both subscales, with medium to large ES (partial $\eta^2 = .09$ and $.15$, respectively). The 2×4 ANOVAs for Total Problems and On-Task revealed significant ($p < .001$) main effects of group, with large ES (partial $\eta^2 = .16$ and $.15$, respectively). There were no significant effects of gender or Gender \times Group interactions for Attention Deficit Hyperactivity Problems, Inattention, Hyperactivity-Impulsivity, Total Problems, or On-Task.

As summarized in Table 2, Tukey HSD tests showed that the ADHD-C group scored significantly higher ($p < .05$) than NON-ADHD REF and Control on the DOF Intrusive and Oppositional syndromes, Attention Deficit Hyperactivity Problems, Hyperactivity-Impulsivity subscale, and Total Problems, and significantly lower than NON-ADHD REF and Control on On-Task. The ADHD-C group also scored significantly higher ($p < .05$) than Control on the Immature/Withdrawn syndrome and Inattention subscale. The ADHD-IN group scored significantly higher ($p < .05$) than Control on the Sluggish Cognitive Tempo and Attention Problems syndromes and Total Problems, and significantly lower than Control on On-Task. The ADHD-C group scored significantly higher ($p < .05$) than ADHD-IN on the Intrusive and Oppositional syndromes and the Hyperactivity-Impulsivity subscale. There were no significant differences between ADHD-IN versus NON-ADHD REF or NON-ADHD REF versus Control on any DOF scale.

SES as a covariate. Because there were significant group differences and site differences on SES, we reran our analyses of DOF scores reported in Table 2, using 2×4 multivariate analyses of covariance (MANCOVAs) and analyses of covariance (ANCOVAs), treating SES as a covariate. Results showed no significant effects of the SES covariate in either MANCOVA (five DOF syndromes or Inattention and Hyperactivity-Impulsivity) or

the ANCOVAs of Attention Deficit Hyperactivity Problems, Total Problems, or On-Task. The SES covariate was significant only in the univariate ANCOVA for the Hyperactivity-Impulsivity subscale, $F(1, 154) = 3.97$, $p = .048$, but this could have been a chance effect (Sakoda, Cohen, & Beall, 1954), and the ES for the SES covariate was small (partial $\eta^2 = .03$).

Age as a covariate. Because there were significant group differences in age, we also re-ran our analyses of DOF scores, using 2×4 MANCOVAs and ANCOVAs, treating age in years as a covariate. These analyses showed no significant effects of the age covariate for any DOF scale.

Discriminant Analyses

ADHD-C versus NON-ADHD REF.

Table 3 shows the cross-validated classification rates for ADHD-C versus NON-ADHD REF derived from various combinations of DOF scales as predictors. The candidate predictors are listed in the first column in the order of their relative contributions to the discriminant functions. When the five DOF syndrome scales were entered as candidate predictors (a), the Intrusive syndrome contributed most to the discriminant function. When the DOF On-Task score was added as a predictor along with the five syndrome scales (b), Intrusive and On-Task, in opposite directions, contributed most to the discriminant function, producing slight improvements in classification rates for ADHD-C and overall correct classification compared to the five syndromes without On-Task. With prior probabilities equal for both groups, the Attention Deficit Hyperactivity Problems scale and On-Task score (c) produced the same classification rates as the five syndromes plus On-Task. The Hyperactivity-Impulsivity and Inattention subscales plus On-Task (d) produced the highest classification rate for ADHD-C (60.9%), with an overall correct classification of 67.8%. When prior probabilities were computed from group sizes, overall correct classification rates

Table 3
Cross-Validated Percentages of Cases Correctly Classified as ADHD-C versus NON-ADHD-REF

DOF Scales Candidate Predictors	Standardized Canonical Coefficients ^a	ADHD-C ^b (<i>n</i> = 64)	NON-ADHD REF ^b (<i>n</i> = 51)	Overall Correct Classification ^b
a. Intrusive	.834	42.2%	78.4%	58.3%
Immature/Withdrawn	.210	(59.4%)	(68.6%)	(63.5%)
Attention Problems	.114			
Sluggish Cognitive Tempo	.105			
Oppositional	.047			
b. Intrusive	-.716	54.7%	70.6%	61.7%
On-Task	.650	(60.9%)	(62.7%)	(61.7%)
Oppositional	-.361			
Immature/Withdrawn	-.222			
Attention Problems	-.035			
Sluggish Cognitive Tempo	-.017			
c. Attention Deficit Hyperactivity Problems	.660	54.7%	70.6%	61.7%
On-Task	-.469	(62.5%)	(60.8%)	(61.7%)
d. Hyperactivity-Impulsivity subscale	-.837	60.9%	76.5%	67.8%
On-Task	.527	(67.2%)	(66.7%)	(67.0%)
Inattention subscale	.311			
e. Total Problems	.667	53.1%	76.5%	63.5%
On-Task	-.453	(65.6%)	(60.8%)	(63.5%)

Note. All discriminant functions were significant at $p \leq .001$. ADHD-C = attention deficit hyperactivity disorder—Combined type; NON-ADHD REF = non-ADHD—clinically referred; DOF = Direct Observation Form.

^aStandardized canonical coefficients from discriminant function analyses with all predictors entered simultaneously.

^bPercentages of cross-validated grouped cases correctly classified with prior probabilities equal for all groups. Percentages in parentheses are cross-validated grouped cases correctly classified with prior probabilities computed from group sizes.

were generally similar to classification rates with prior probabilities equal for both groups.

ADHD-IN versus NON-ADHD REF. Discriminant analyses for ADHD-IN versus NON-ADHD REF produced significant discriminant functions ($p < .05$) only for the five DOF syndrome scales alone as predictors and for the five syndrome scales plus the On-Task score as predictors. Among the five syndromes, Attention Problems and Sluggish Cognitive Tempo, plus Oppositional in reverse direction, contributed most to the discriminant function. When the On-Task score was added to the five syndromes as predictors, Oppositional, Attention Problems, and Sluggish Cognitive Tempo continued to contribute most to the discriminant function, followed by On-Task (standardized ca-

nonical coefficients = $-.675$, $.594$, $.493$, and $-.424$, respectively). With prior probabilities equal for both groups, classification rates were 45.5% for ADHD-IN and 70.6%–74.5% for NON-ADHD REF, with overall correct classification rates of 63–65.8%. With prior probabilities computed from group sizes, much higher classification rates were obtained for NON-ADHD REF (86.8%–88.2%) versus lower classification rates for ADHD-IN (18.2%–22.7%), with overall correct classification of 67.1%.

ADHD-IN versus Control. Table 4 shows the cross-validated classification rates for ADHD-IN versus Control. When the five DOF syndrome scales were entered as candidate predictors (a), the Attention Problems and Sluggish Cognitive Tempo syndromes con-

Table 4
Cross-Validated Percentages of Cases Correctly Classified as ADHD-IN
versus Control

DOF Scales Candidate Predictors	Standardized Canonical Coefficients ^a	ADHD-IN ^b (<i>n</i> = 22)	Control ^b (<i>n</i> = 26)	Overall Correct Classification ^b
a. Attention Problems	.606	68.2%	73.1%	70.8%
Sluggish Cognitive Tempo	.583	(68.2%)	(80.8%)	(75.0%)
Immature/Withdrawn	.314			
Intrusive	-.285			
Oppositional	.002			
b. On-Task	.627	68.2%	73.1%	70.8%
Attention Problems	-.554	(54.5%)	(80.8%)	(68.8%)
Sluggish Cognitive Tempo	-.414			
Intrusive	.350			
Oppositional	.303			
Immature/Withdrawn	-.248			
c. On-Task	.721	72.7%	69.2%	70.8%
Attention Deficit Hyperactivity Problems	-.429	(72.7%)	(73.1%)	(72.9%)
d. On-Task	.634	54.5%	69.2%	62.5%
Inattention subscale	-.533	(54.5%)	(73.1%)	(64.6%)
Hyperactivity-Impulsivity subscale	.011			
e. On-Task	.686	63.6%	73.1%	68.8%
Total Problems	-.479	(63.6%)	(73.1%)	(68.8%)

Note. All discriminant functions were significant at $p < .01$, except (a), which was $p = .02$. ADHD-IN = attention deficit hyperactivity disorder—Inattentive type; Control = typically developing controls; DOF = Direct Observation Form.

^aStandardized canonical coefficients from discriminant function analyses with all predictors entered simultaneously.

^bPercentages of cross-validated grouped cases correctly classified with prior probabilities equal for all groups. Percentages in parentheses are cross-validated grouped cases correctly classified with prior probabilities computed from group sizes.

tributed most to the discriminant function. When the DOF On-Task score was added to the five syndromes as predictors (b), it contributed most to the discriminant function, followed by Attention Problems and Sluggish Cognitive Tempo. The DOF On-Task score also contributed the most to the discriminant functions in the remaining combinations of DOF scales (c)–(e). When prior probabilities were equal for both groups, the On-Task score with the Attention Deficit Hyperactivity Problems scale (c) produced the highest classification rate for ADHD-IN (72.7%), but the overall correct classification of 70.8% was the same as for the five syndromes with and without On-Task. Somewhat lower classification

rates were obtained for combinations of On-Task with the Inattention and Hyperactivity-Impulsivity subscales and On-Task with Total Problems (d) and (e). When prior probabilities were computed from group sizes, there were only slight changes in overall correct classification rates.

ADHD-C versus Control. Discriminant analyses for ADHD-C versus Controls showed that among the five DOF syndrome scales, Sluggish Cognitive Tempo, Intrusive, and Oppositional contributed most to the discriminant function. When the On-Task score was added as a predictor with the five syndromes, it contributed the most to the dis-

criminant function, followed by the Intrusive and Sluggish Cognitive Tempo syndromes (standardized canonical coefficients = .771, $-.336$, and $-.273$, respectively). The On-Task score also contributed most to discriminant functions in the remaining combinations of DOF scales, similar to results for ADHD-IN versus Control. With prior probabilities equal for both groups, classification rates ranged from 57.8% to 64.1% for ADHD-C and 76.9% to 80.8% for Control, with overall correct classification rates of 63.3%–67.8%. With prior probabilities computed from group sizes, much higher classification rates were obtained for ADHD-C (87.5%–95.3%) versus lower classification rates for Control (23.1%–53.8%), with overall correct classification of 74.4%–83.3%.

ADHD-C versus ADHD-IN. Discriminant analyses for ADHD-C versus ADHD-IN produced significant discriminant functions ($p < .05$) for the five DOF syndrome scales alone, the five syndrome scales plus On-Task, and the Inattention and Hyperactivity-Impulsivity subscales plus On-Task as predictors. Among the five syndromes, Attention Problems, Intrusive, and Oppositional contributed most to the discriminant function (standardized canonical coefficients = $-.664$, $.652$, and $.493$, respectively). When the On-Task score was added to the five syndromes as predictors, Attention Problems, Intrusive, and On-Task contributed most to the discriminant function, followed by Sluggish Cognitive Tempo (standardized canonical coefficients = $.655$, $-.603$, $.417$, and $.382$, respectively). DOF Hyperactivity-Impulsivity, Inattention and On-Task had similarly high contributions to the discriminant function (standardized canonical coefficients = $-.832$, $.743$, and $.737$, respectively). When prior probabilities were equal for both groups, classification rates were 57.8%–67.2% for ADHD-C and 59.1%–68.2% for ADHD-IN, with overall correct classification rates of 60.5%–67.4%. However, with prior probabilities computed from group sizes, classification rates were 95.3%–100% for ADHD-C and 0%–18.2% for ADHD-IN, with overall correct classification of 74.4%–75.6%.

Discussion

Direct observation of children's behavior in school classrooms can be an important adjunct to parent and teacher reports for assessing behaviors consistent with ADHD. In the present study, we used a standardized form, the DOF, to examine group differences in observations of classroom behavior for 6–11-year-old children with DSM-IV-TR diagnoses of the ADHD-C and ADHD-IN subtypes versus other clinically referred children without ADHD and nonreferred control children. In a parallel study, McConaughy et al. (2009) used a similar Achenbach System of Empirically Based Assessment form, the Test Observation Form (TOF; McConaughy & Achenbach, 2004), to examine differences in test session observations for the same four diagnostic groups drawn from the same sample. Comparisons of findings across the two studies highlight consistencies and inconsistencies in behavior observed in school classrooms versus behavior observed by examiners in test sessions.

ADHD-C versus NON-ADHD REF and Control

Consistent with our hypotheses, on average, children with ADHD-C scored significantly higher than NON-ADHD REF and Control on the DOF DSM-oriented Attention Deficit Hyperactivity Problems scale and Hyperactivity-Impulsivity subscale, and significantly lower on DOF On-Task, with large ES (14–15% of variance). Our hypothesis regarding the DOF Inattention subscale was partially supported with the ADHD-C group scoring significantly higher than Control (9% of variance), but not higher than NON-ADHD REF. Discriminant analyses combining the DOF problem scales with On-Task scores as predictors produced overall correct classification rates of 61.7%–67.8% for ADHD-C versus NON-ADHD REF and 63.3%–67.8% for ADHD-C versus Control.

Our results are consistent with findings reported by McConaughy et al. (2009) for the TOF Attention Deficit Hyperactivity Problems scale, Hyperactivity-Impulsivity subscale, and

Inattention subscale, except in that study ADHD-C also scored significantly higher than NON-ADHD REF on TOF Inattention, with medium to large ES (15%–20% of variance). The similar findings across the two studies demonstrate considerable consistency between observations of classroom behavior and observations of test session behavior in children from the same sample. Similar findings across the two settings are even more remarkable considering that the raters were different (test examiners versus classroom observers) and both sets of raters were blind to all assessment results and children's diagnostic group assignments. Significant differences between the ADHD-C group versus Control on the DOF and TOF Inattention subscales are also consistent with the findings of Glutting, Robins, and deLancy (1997) for the Inattention scale of the Guide to Assessment of Test Session Behavior. Our findings for DOF On-Task scores are consistent with several other observational studies of on-task and off-task behavior of children with ADHD versus typically developing peers in the same classrooms (DuPaul, Volpe, Jitendra, Lutz, Lorah, and Gruber, 2004; Gadow et al., 1996; Junod, DuPaul, Jitendra, Volpe, & Cleary, 2006; Skansgaard & Burns, 1998).

The ADHD-C group also scored significantly higher than NON-ADHD REF and Control on the DOF Oppositional syndrome, consistent with our hypotheses, as well as the DOF Intrusive syndrome, with medium to large ES (9%–13% of variance). These results are consistent with findings of McConaughy et al. (2009) for the TOF Oppositional syndrome and findings of Skansgaard and Burns (1998) for their ODD/covert CD scale created from the 1986 DOF. From parent reports on the NIMH DISC-4, we also found higher rates of comorbid DSM-IV-TR diagnoses of CD and ODD for children with ADHD-C than for the NON-ADHD REF and Control groups. In the Multimodal Treatment Study of Children with Attention Deficit Hyperactivity Disorder (MTA), Abikoff et al. (2002) also reported significantly higher rates of observed rule-breaking and impulsive and aggressive behav-

ior for children with ADHD-C who had comorbid disruptive disorders versus children with comorbid anxiety disorders or no comorbid disorders. Analyzing effects of comorbid diagnoses on DOF scores was beyond the scope of our study and, according to power analyses, not feasible given our sample sizes. However, the high rates of comorbidity can have important clinical implications, because researchers in the MTA study reported significant differences in functional impairment and response to treatment for children with ADHD who also have comorbid disruptive disorders and/or anxiety disorders (Jensen et al., 2001).

Contrary to our hypotheses, there were no significant differences between ADHD-C and NON-ADHD REF or Control on the DOF Attention Problems syndrome. Discriminant analyses also showed that DOF Attention Problems was a minimal contributor to differentiating ADHD-C from NON-ADHD REF or Control. The null findings for DOF Attention Problems are inconsistent with McConaughy et al.'s (2009) findings for test session observations, where the TOF Attention Problems syndrome was a strong predictor of ADHD-C versus the other two groups without ADHD. The differences in findings across the two studies may be partially from the larger number of problem items comprising TOF Attention Problems (17) compared to DOF Attention Problems (8), as well as differences between types of observers and setting (one-on-one test sessions versus classroom).

ADHD-IN versus Control and NON-ADHD REF

Consistent with our hypotheses, children with ADHD-IN scored significantly higher than Control on the DOF Sluggish Cognitive Tempo and Attention Problems syndromes and the Inattention subscale, and lower on DOF On-Task, with medium to large ES (8%–15% of variance). In discriminant analyses, DOF Attention Problems and Sluggish Cognitive Tempo contributed most among the five syndromes for predicting ADHD-IN versus Control, with an overall correct classification

of 70.8%. The addition of the On-Task score to the five syndromes produced the same classification rates. These results are consistent with findings reported by Skansgaard and Burns (1998) for their Slow Cognitive Tempo and Inattention scales created from the 1986 DOF. By contrast, McConaughy et al. (2009) found no differences between children with ADHD-IN versus Control on any TOF scales, including TOF Attention Problems and TOF Inattention subscale. The contrast in findings from the three studies suggests that observations of attention problems and sluggish cognitive tempo in the classroom may be better than observations in test sessions for differentiating children with ADHD-IN from typically developing children. Perhaps this is because children with ADHD-IN receive less reinforcement for attending in classroom settings than in one-on-one test sessions, although this needs to be tested in future research.

Contrary to our hypotheses, we found no significant differences between ADHD-IN versus NON-ADHD REF on any DOF scale, similar to findings by McConaughy et al. (2009) for the TOF scales. The null findings from both studies suggest that behavioral observations in either setting are not valid methods alone for differentiating children with ADHD-IN from children with other clinical problems. McConaughy et al. (2009) did report significantly lower intelligence and achievement test scores for children with ADHD-IN (and ADHD-C) versus NON-ADHD REF and Control. However, even the lower test scores for ADHD-IN must be interpreted with caution because some of the significant effects disappeared when learning disability was entered as a covariate.

ADHD-C versus ADHD-IN subtypes

Differentiating between the DSM-IV-TR ADHD-C and ADHD-IN subtypes is a more challenging clinical task than differentiating each subtype from typically developing children or other clinically referred children without ADHD. Consistent with our hypotheses, we found that the ADHD-C group scored significantly higher than ADHD-IN on the

DOF Hyperactivity-Impulsivity subscale and Oppositional syndrome. The ADHD-C group also scored significantly higher than ADHD-IN on the DOF Intrusive syndrome. Our results are consistent with McConaughy et al.'s (2009) findings for the TOF Oppositional syndrome and Hyperactivity-Impulsivity subscale and Skansgaard and Burns' (1998) findings for their Hyperactivity-Impulsivity and ODD/overt CD scales created from the 1986 DOF.

Contrary to our hypotheses, we found no significant differences between the ADHD-C versus ADHD-IN groups on the DOF Attention Problems syndrome or the Attention Deficit Hyperactivity Problems scale, both of which included attention problems along with hyperactivity and impulsivity. These null results are inconsistent with McConaughy et al.'s (2009) findings, wherein the ADHD-C group scored significantly higher than ADHD-IN on TOF Attention Problems and Attention Deficit Hyperactivity Problems, as well as the TOF Inattention subscale. The contrast in findings across the two studies suggests that both setting and the type of observed problem behaviors are important to consider in attempts to differentiate the ADHD-C and ADHD-IN subtypes through observational measures. Specifically, similar findings for the DOF and TOF Oppositional syndrome and Hyperactivity-Impulsivity subscale suggest that children with ADHD-C are likely to exhibit higher levels of these kinds of problems in both settings than will children with ADHD-IN. At the same time, differences in findings for the DOF and TOF Attention Problems syndromes, Attention Deficit Hyperactivity Problems scales, and Inattention subscale suggest that children with ADHD-C are likely to display more severe attention problems than children with ADHD-IN in one-on-one test sessions, but not in school classrooms. It was also notable that McConaughy et al. (2009) found no significant differences between the two ADHD subtypes on intelligence or achievement test scores, consistent with other studies (e.g., Solanto et al., 2007).

Limitations

There are several limitations to our study. First, sample sizes were relatively small for ADHD-IN and Control compared to the other two groups. The small sample size for ADHD-IN, in particular, could have reduced power for finding differences between this group and ADHD-C and NON-ADHD REF. Power analyses also indicated that our samples were too small to allow secondary analyses of cases with comorbid DSM-IV-TR diagnoses. A second limitation was that our sample included only 6- to 11-year-old children, so the results may not generalize to adolescents. A third limitation was that classroom observers may have developed some hypotheses about the children that could have affected their ratings on the DOF. To minimize rater bias, observers were kept blind to all clinical information about the children and they were provided detailed behavioral descriptors as guidelines for scoring the DOF problem items and On-Task. A fourth limitation was that ratings on each DOF were based only on 10-min samples of behavior. However, by obtaining four DOFs for the vast majority of our sample, we were able to derive problem scores and on-task ratings of behavior observed for a total of 40 min, including mornings and afternoons of two different days.

Conclusions and Practical Implications

In summary, medium to large group effects and good classification rates in the present study provided strong evidence of the discriminative validity of the DOF Attention Deficit Hyperactivity Problems scale, Hyperactivity-Impulsivity subscale, and the DOF Oppositional and Intrusive syndromes, for differentiating children with the ADHD-C subtype from other children without ADHD. The DOF Hyperactivity-Impulsivity subscale and Oppositional and Intrusive syndromes also significantly differentiated between the ADHD-C and ADHD-IN subtypes. Furthermore, medium to large group effects and good classification rates supported the discriminative validity of the DOF Sluggish Cognitive Tempo and Attention Problems syndromes,

and the Inattention subscale, for differentiating children with ADHD-IN subtype from typically developing control children. The DOF On-Task score was also a good discriminator of both ADHD subtypes versus typically developing controls.

The findings from this study have important practical implications for school psychologists. School psychologists often conduct direct observations to assess behavior problems, including ADHD (Shapiro & Heick, 2004; Demaray et al., 2003). The DOF provides a standardized and efficient method for recording and quantifying such observations. School psychologists can use the DOF in several ways. First, the DOF can be used to screen for problem behaviors that warrant further assessment of potential ADHD. School psychologists should pay special attention to high scores on the DOF scales summarized in the preceding paragraph that showed good discriminative validity for differentiating between children with and without ADHD and differentiating between the ADHD subtypes. Second, high scores on relevant DOF scales can be used to corroborate parent and teacher reports in comprehensive assessments of ADHD. Considering the moderate agreement usually found between parents and teachers (Achenbach et al., 1987; Mitsis et al., 2000), DOF scale scores may be especially useful for diagnostic decision making for cases in which parents and teachers disagree in their reports of children's problems. In addition, the present findings on the DOF showed that many children with ADHD-C exhibit multiple problems, including oppositional and intrusive behavior, along with ADHD-consistent behaviors. School psychologists need to keep this multiplicity of problems in mind when evaluating children referred for disruptive behavior in their classrooms. Finally, school psychologists can use the DOF to evaluate and monitor effects of interventions in formative assessments. In a separate study, Volpe et al. (2009) reported strong generalizability and dependability for the DOF Attention Deficit Hyperactivity Problems scale, Hyperactivity-Impulsivity subscale, and Oppositional and Intrusive syndromes, used over multiple observation

sessions. Volpe et al. also reported that these same four DOF scales, plus Total Problems, required the fewest number of 10-min observations to reach acceptable reliability. Volpe et al.'s findings, combined with findings of the present study, support the discriminative validity and practical utility of the DOF for assessing ADHD. At the same time, standardized observational data from the DOF must be combined with information from other sources, including parent and teacher reports, for making diagnostic decisions and planning interventions.

Footnotes

¹ We computed alphas and Pearson r to assess reliability of the DOF scales consistent with procedures to assess reliability of other Achenbach System of Empirically Based Assessment scales (Achenbach & Rescorla, 2001).

² To adjust for unequal sample sizes, the Tukey HSD test on homogenous subsets uses the harmonic mean sample size and $\alpha = .05$ to test group differences. However, multiple comparisons based on MANOVAs and univariate ANOVAs in our analyses showed most significant group differences at $p < .01$ or lower.

References

- Abikoff, H. B., & Gittelman, R. (1985). Classroom Observation Code: A modification of the Stony Brook Code. *Psychopharmacology Bulletin*, *21*, 901–909.
- Abikoff, H. H., Jensen, P. S., Arnold, L. L. E., Hoza, B., Hechtman, L., Pollack, S., et al. (2002). Observed classroom behavior of children with ADHD: Relationship to gender and comorbidity. *Journal of Abnormal Child Psychology*, *30*, 349–359.
- Achenbach, T. M. (1986). *The Direct Observation Form of the Child Behavior Checklist* (rev. ed.). Burlington: University of Vermont, Department of Psychiatry.
- Achenbach, T. M., McConaughy, S. H., & Howell, C. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213–232.
- Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington: University of Vermont, Research Center for Children, Youth, and Families.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.
- Barkley, R. A. (2006). *Attention deficit hyperactivity disorder: A handbook for diagnosis and treatment* (3rd ed.). New York: Guilford Press.
- Browne, N. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.
- Demaray, M. K., Schaefer, K., & Delong, L. K. (2003). Attention-Deficit/Hyperactivity Disorder (ADHD): A national survey of training and current assessment practices in the schools. *Psychology in the Schools*, *40*, 583–597.
- DuPaul, G. J., Power, T. J., Anastopoulos, A. D., & Reid, R. (1998). *Manual for the ADHD Rating Scale—IV*. New York: Guilford Press.
- DuPaul, G. J., & Stoner, G. (2003). *ADHD in the schools* (2nd ed.) New York: Guilford Press.
- DuPaul, G. J., Volpe, R. J., Jitendra, A. K., Lutz, J. G., Lorah, K. S., & Grubner, R. (2004). Elementary school students with attention-deficit/hyperactivity disorder: Predictors of academic achievement. *Journal of School Psychology*, *42*, 285–301.
- Gadow, K. D., Sprafkin, J., & Nolan, E. E. (1996). *ADHD School Observation Code*. Stony Brook, NY: Checkmate Plus.
- Glutting, J. J., Robins, P. M., & de Lancy, E. (1997). Discriminant validity of test observations for children with attention deficit hyperactivity disorder. *Journal of School Psychology*, *35*, 391–401.
- Hollingshead, A. B. (1975). *Four factor index of social status*. Unpublished paper. New Haven, CT: Yale University, Department of Sociology.
- Jensen, P. T., Hinshaw, S. P., Kraemer, H. C., Lenora, N., Newcorn, J. H., Abikoff, H. B., et al. (2001). ADHD comorbidity findings from the MTA Study: Comparing comorbid subtypes. *Journal of the American Academy of Child and Adolescent Psychiatry*, *40*, 147–158.
- Junod, R. E. V., DuPaul, G. J., Jitendra, A. S., Volpe, R. J., & Cleary, K. S. (2006). Classroom observations of students with and without ADHD: Differences across types of engagement. *Journal of School Psychology*, *44*, 87–104.
- McConaughy, S. H., & Achenbach, T. M. (2004). *Manual for the ASEBA Test Observation Form*. Burlington: University of Vermont, Research Center for Children, Youth, & Families.
- McConaughy, S. H., & Achenbach, T. M. (2009). *Manual for the ASEBA Direct Observation Form*. Burlington: University of Vermont, Research Center for Children, Youth, & Families.
- McConaughy, S. H., Achenbach, T. M., & Gent, C. L. (1988). Multiaxial empirically-based assessment: Parent, teacher, observational, cognitive, and personality correlates of Child Behavior Profiles for 6–11 year-old boys. *Journal of Abnormal Child Psychology*, *16*, 485–509.
- McConaughy, S. H., Kay, P. J., & Fitzgerald, M. (1999). The Achieving, Behaving, Caring Project for preventing ED: Two-year outcomes. *Journal of Emotional and Behavioral Disorders*, *7*, 224–239.
- McConaughy, S. H., Ivanova, M., Antshel, K., & Eiraldi, R. B. (2009). Standardized observational assessment of attention deficit hyperactivity disorder combined and

- predominantly inattentive subtypes: I. Test session observations. *School Psychology Review*, 38, 45–66.
- 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.
- Nigg, J. T. (2006). *What causes ADHD?* New York: Guilford Press.
- Platzman, K. A., Stoy, M. R., Brown, R. T., Coles, C., Smith, I. E., & Falek, A. (1992). Review of observational methods in attention deficit hyperactivity disorder (ADHD): Implications for diagnosis. *School Psychology Quarterly*, 7, 155–177.
- Power, T. J., Andrews, T. J., Eiraldi, R. B., Doherty, B. J., Ikeda, M. J., DuPaul, G. J., et al. (1998). Evaluating attention deficit hyperactivity disorder using multiple informants: The incremental utility of combining teacher with parent reports. *Psychological Assessment*, 10, 250–260.
- Reed, M. L., & Edelbrock, C. (1983). Reliability and validity of the Direct Observation Form of the Child Behavior Checklist. *Journal of Abnormal Child Psychology*, 11, 521–530.
- Sakoda, J. M., Cohen, B. H., & Beall, G. (1954). Test of significance for a series of statistical tests. *Psychological Bulletin*, 51, 172–175.
- Shaffer, D., Fisher, P., Lucas, C. P., Dulcan, M., & Schwab-Stone, M. E. (2000). 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*, 39, 28–38.
- Shapiro, E. S. (2004). *Academic skills problems workbook* (rev.). New York: Guilford Press.
- Shapiro, E. S., & Heick, P. (2004). School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. *Psychology in the Schools*, 41, 551–561.
- Skansgaard, E. P., & Burns, G. L. (1998). 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*, 20, 1–14.
- Solanto, M. V., Gilbert, S. N., Raj, A., Zhu, J., Pope-Boyd, S., Stepak, B., et al. (2007). Neurocognitive functioning in AD/HD, predominantly inattentive and combined subtypes. *Journal of Abnormal Child Psychology*, 35, 729–744.
- SPSS. (2007). *SPSS Base 15.0 user's guide*. Chicago, IL: Author.
- Volpe, R. J., McConaughy, S. H., & Hintze, J. (2009). Generalizability of classroom behavior problem and on-task scores from the Direct Observation Form. *School Psychology Review*, 38, 382–401.
- Volpe, R., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in classroom settings : A review of seven coding schemes. *School Psychology Review*, 34, 454–474.
- Wilson, M. S., & Reschly, D. J. (1996). Assessment in school psychology training and practice. *School Psychology Review*, 25, 9–23.
- Yu, C. Y., & Muthen, B. O. (2002). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes* (Technical report). Los Angeles: University of California at Low Angeles, Graduate School of Education and Information Studies.

Stephanie H. McConaughy, PhD, is Research Professor of Psychiatry and Psychology Emerita at the University of Vermont. She is a Vermont-licensed practicing psychologist and nationally certified school psychologist. Her research focuses on empirically based assessment of children's behavioral and emotional problems, multimethod assessment of ADHD, and school-based prevention programs for behavioral disorders.

Masha Y. Ivanova is Research Assistant Professor of Psychiatry at the University of Vermont. She received her PhD at the University of Albany and completed her predoctoral clinical internship and postdoctoral training at the University of Vermont Center for Children, Youth, and Families. Her research focuses on the understanding of environmental factors as risk and protective factors for child psychopathology.

Kevin Antshel, PhD, is Assistant Professor of Psychiatry and Director of the Adult ADHD Treatment & Research Program at the State University of New York (SUNY)-Upstate Medical University. His research and clinical interests include ADHD, learning disabilities, and developmental neuropsychology.

Ricardo B. Eiraldi, PhD, is Assistant Professor of Clinical Psychology in the Department of Pediatrics at the University of Pennsylvania. His research focuses on the clinical presentation of ADHD in girls and ethnic minority populations; the application of help-seeking behavior models in the study of health disparities among ethnic minority children and families; and the development of strategies for addressing mental health services disparities in the inner city.

Levent Dumenci, PhD, is Associate Professor of Social and Behavioral Health at Virginia Commonwealth University (VCU) and Director of Behavioral Measurement Core at the VCU Massey Cancer Center. He is a doctoral-trained psychometrician and statistician. His research focuses on latent mixture extensions of traditional measurement models involving continuous and discrete latent variables, measurement of change, and test development.

Date Received: April 7, 2009

Date Accepted: June 29, 2009

Action Editor: George Bear ■