Journal of Abnormal Child Psychology, Vol. 27, No. 6, 1999, pp. 463–472

Utility of Behavior Ratings by Examiners During Assessments of Preschool Children With Attention-Deficit/Hyperactivity Disorder

Erik G. Willcutt,[1,6] Cynthia M. Hartung,[2] Benjamin B. Lahey,[3] Jan Loney,[4] and William E. Pelham[5]

Received February 4, 1999; revision received May 26, 1999; accepted May 27, 1999

This study examines the clinical utility of behavior ratings made by nonclinician examiners during assessments of preschool children with Attention-Deficit/Hyperactivity Disorder (AD/HD). Matched samples of children with ($n = 127$) and without ($n = 125$) AD/HD were utilized to test the internal, convergent, concurrent, and incremental validity of ratings completed by examiners on the Hillside Behavior Rating Scale (HBRS). Results indicated that HBRS ratings were internally consistent, possessed sufficient interrater reliability, and were significantly associated with parent and teacher reports of AD/HD when controlling for age, gender, intelligence, and symptoms of other psychopathology. HBRS ratings also were significantly associated with other measures of functioning, and provided a significant increment in the prediction of impairment over parent and teacher report alone. These findings suggest that behavioral ratings during testing provide a unique source of clinical information that may be useful as a supplement to parent and teacher reports.

Numerous previous studies have examined the reliability and validity of parent and teacher reports of Attention-Deficit/Hyperactivity Disorder (AD/HD) symptoms (Barkley, 1988; Barkley & Edelbrock, 1987; DuPaul, 1991; DuPaul & Barkley, 1992; DuPaul & Stoner, 1994; Hart, Lahey, Loeber, & Hanson, 1994; Hinshaw, 1994; Lahey *et al.*, 1994, 1998, 1999; Molina, Pelham, Blumenthal, & Galiszewski, 1998; Pelham, Gnagy, Greenslade, & Milich, 1992). Results of these studies suggest that reports by parents and teachers are reliable and validly identify children with AD/HD who exhibit significant functional impairment. In contrast, few studies have examined the diagnostic utility of ratings of children's behavior during testing as part of a comprehensive diagnostic assessment of AD/HD.

Sleator and Ullmann (1981) suggest that many children with AD/HD may not exhibit significant symptoms in the clinic environment, possibly because of the novelty of the setting or the anxiety aroused by the experience. This hypothesis suggests that the inclusion of observable inattentive, hyperactive, or impulsive behaviors in the clinic as a criterion for diagnosis would precipitate a high rate of incorrect diagnoses. In the only other systematic study of this issue, Barkley (1998) reported preliminary findings suggesting that neither negative nor positive clinician ratings of a child's AD/HD behaviors during an evaluation were significantly associated with symptoms of AD/HD or Oppositional Defiant Disorder (ODD) reported by parents. In contrast, 70% of children who exhibited significant AD/HD symptomatology in the clinic setting were also rated as significantly elevated by teachers.

The present study examined the utility of behavior ratings on the Hillside Behavior Rating Scale (HBRS; Gittelman & Klein, 1985) as part of a clinical assessment of AD/HD in preschoolers. To have diagnostic utility

[1]University of Colorado at Boulder, Boulder, Colorado 80309.
[2]Oregon Health Sciences University, Portland, Oregon.
[3]University of Chicago, Chicago, Illinois.
[4]State University of New York at Stony Brook, Stony Brook, New York.
[5]State University of New York at Buffalo, Buffalo, New York.
[6]Address correspondence to Erik Willcutt, Institute for Behavioral Genetics, University of Colorado at Boulder, Campus Box 447, Boulder, Colorado 80309.

0091-0627/99/1200-0463$16.00/0

as part of an assessment, clinic ratings must have sufficient reliability, must be significantly associated with other measures of AD/HD, and must provide a significant increment in the prediction of functional impairment in domains that are not assessed by the symptoms of AD/HD. Four methods were utilized to test if HBRS ratings provided a useful addition to parent and teacher reports as part of a comprehensive diagnostic assessment. First, the internal consistency and interrater reliability of HBRS ratings were examined to determine whether these ratings were sufficiently reliable to be useful. Next, the convergent validity of the ratings was evaluated by testing if clinic ratings of AD/HD behaviors were significantly associated with parent and teacher reports of AD/HD symptomatology. Third, the concurrent validity of HBRS ratings was assessed by examining the relation with measures of functional impairment. Finally, the incremental validity of HBRS ratings was assessed by determining if these ratings provided a significant improvement in the prediction of functional impairment beyond the prediction provided by parent and teacher reports.

METHOD

Participants

Participants included 127 preschool children with DSM-IV [American Psychiatric Association (APA), 1994] AD/HD (22 females, 105 males) and 125 children without AD/HD (24 females, 101 males). The diagnosis of AD/HD was based on the "or-rule," which codes each AD/HD symptom as positive if it is endorsed by either the child's parent or teacher (Piacentini, Cohen, & Cohen, 1992). Other analyses of data from the present sample indicate that this procedure provides the optimal combination of diagnostic specificity and sensitivity when functional impairment is utilized as the standard for accurate classification (Lahey *et al.*, 1999).

A detailed description of the inclusion and exclusion criteria for the study is provided by Lahey *et al.* (1998). Participants ranged from 3 years 10 months to 7 years 0 months of age at the time of the assessment, with 98.4% being 4 through 6 years old at their last birthday. Participants with AD/HD were recruited in Chicago ($n = 58$) through a university child psychiatry clinic, and were recruited in Pittsburgh ($n = 69$) either through a university child psychiatry clinic (42%) or through newspaper advertisements and flyers distributed to schools. Participants were referred to the study by the clinics if they lived with their biological mother and did not exhibit pervasive developmental disorder, psychosis, or clear neurological disorder. Site was entered as a covariate in all initial models to control for any differences in recruiting strategies, but was dropped from the final models because it had no significant impact on any result.

Participants with and without AD/HD were matched on ethnicity, age, gender, and family income. The ethnic composition of the full sample was 64% non-Hispanic White, 30% African American, and 6% other ethnic origin. Both groups had a mean age of 5.2 years, and the mean yearly family income was not significantly different between the two groups (AD/HD $M = \$38{,}701$; Non-ADHD $M = \$44{,}959$), $t(250) = 1.86$, $p > .05$. However, participants with AD/HD had significantly lower scores on the Stanford-Binet-IV intelligence composite ($M = 92.0$) than participants without AD/HD ($M = 101.3$), $t(250) = 5.42$, $p < .001$.

Procedures

Assessments were completed during a single visit to the clinic, with information obtained from teachers after the clinic assessment. All measures were administered by two trained nonclinician interviewers, each of whom had at least a bachelor's degree in psychology, social work, or education, and experience working with children. Both examiners were unaware whether the child was in the AD/HD group or the comparison sample. One examiner (hereafter referred to as the "interviewer") interviewed the parent while the other (the "tester") administered the tests of intelligence and academic achievement. After completing the parent interview, the interviewer administered a questionnaire about friendships to the child. The interviewer and the tester were unaware of the results of the measures collected by the other examiner. At the end of the assessment session, but prior to any discussion with one another, the interviewer and the tester each completed the HBRS and the interviewer completed the Children's Global Assessment Scale (CGAS). Raters were instructed to base their HBRS ratings on the child's behavior during the administration of the measures.

Measures

Ratings of Clinic Behavior

The HBRS is a seven-item scale designed for observers to rate a child's behavior. Three HBRS items assess domains that are directly analogous to symptoms of DSM-IV AD/HD (motor activity, distractibility, impulse control) and four items assess more general disruptive behaviors (frustration tolerance, cooperation, interest in tasks, attention seeking). In contrast to the Likert-scale

Table I. Sample Rating Scale for Gross Motor Activity Item from the Hillside Behavior Rating Scale

Score	Description
1.	*Average in gross motor activity.* (No excessive running, no restlessness, or fidgety behavior)
2.	*Somewhat restless.* May have movements of the hands, fingers, or arms. Does not have excessive gross motor activity such as running, climbing, inability to sit still.
3.	*Restless and fidgety.* Can sit for appropriate periods of time, but squirms in chair, moves about in the chair, does not sit still.
4.	*Very restless and fidgety.*
5.	*Hyperactive, but can be controlled.* Has difficulty sitting down. Gets up, but can be brought back.
6.	*Very hyperactive, very difficult to control.* Difficult to get back.
7.	*Severe hyperactivity.* Cannot be controlled; runs around, climbs, cannot sit for any length of time.

format of most rating scales commonly utilized to assess AD/HD, the HBRS utilizes specific operational definitions of behavior as anchor points within each domain (see Table I for a sample item from the HBRS). The rater is asked to select one of five to seven descriptors that best describes the child's behavior during the observation period.

Results of previous studies suggest that HBRS ratings by observers in the classroom have satisfactory reliability and discriminant validity, and are sensitive to the effects of stimulant medication treatment (Abikoff & Gittelman, 1985; Klein & Abikoff, 1997). Specifically, reliability coefficients for the HBRS items ranged from .68 to .76 (Abikoff & Gittelman, 1985), and all items except frustration tolerance significantly discriminated children with AD/HD from control children prior to treatment. HBRS ratings of children with AD/HD improved significantly after 4 weeks of treatment with stimulant medication, with the most pronounced change on items measuring motor activity and impulse control.

Because the goal of the current study was to evaluate the validity of ratings of specific AD/HD behaviors during testing, a composite HBRS AD/HD score was computed by summing the three HBRS items that correspond directly to the dimensions of DSM-IV AD/HD. This AD/HD composite was utilized for all analyses described in this report. To test if the utilization of a subset of the HBRS items significantly influenced the results, analyses were repeated with a composite score comprising the sum of the seven HBRS items. The pattern of results was virtually identical whether the AD/HD composite or HBRS total score was utilized, a finding that is not surprising in light of the extremely high correlation between the composite and the total score ($r = .96$).

Diagnostic Measures

The National Institute of Mental Health Diagnostic Interview Schedule for Children, Version 2.3 (DISC-2.3; Shaffer, Fisher, Piacentini, Schwab-Stone, & Wicks, 1993) was administered to the biological mother to assess symptoms of DSM-III-R AD/HD (APA, 1987). A supplementary module from the DSM-IV field trials (Lahey *et al.*, 1994) was also administered to the mother to assess symptoms of DSM-IV AD/HD that were not included in DSM-III-R. Previous studies (e.g., Schwab-Stone *et al.*, 1996) have shown that test–retest agreement is good to excellent for AD/HD as assessed by the DISC-2.3 ($\kappa = .65$–$.80$). Symptoms of AD/HD in the school setting were assessed by asking the primary classroom teacher of each participant to complete the DSM-IV version of the Disruptive Behavior Disorder (DBD) checklist (Pelham *et al.*, 1992). Consistent with previous studies (e.g., Milich, Hartung, Martin, & Haigler, 1994; Pelham *et al.*, 1992), items rated as "pretty much" or "very much" were scored as positive symptoms. The internal reliability of AD/HD in this sample was high for reports by both parents ($\alpha = .96$) and teachers ($\alpha = .96$).

Comorbid Psychopathology

The DISC-2.3 was also utilized to obtain maternal reports of symptoms of ODD, Conduct Disorder (CD), anxiety disorders, and mood disorders. Schwab-Stone *et al.* (1996) reported that test–retest agreement is moderate to good for ODD and CD ($\kappa = .56$–$.73$) and moderate for anxiety and depressive disorders ($\kappa = .50$–$.64$). Teacher reports of ODD and CD symptoms were obtained from the DBD checklist. Estimates of internal reliability in the present sample were acceptable for all diagnoses ($\alpha = .73$–$.93$), a finding that is consistent with previous studies of these measures (e.g., Pelham *et al.*, 1992; Shaffer *et al.*, 1996).

Intelligence and Academic Readiness/Achievement

Intelligence was estimated by the standard Short Form of the Stanford-Binet Intelligence Scale (4th ed.; Thorndike, Hagan, & Sattler, 1986). Academic readiness/achievement in reading and mathematics was assessed by the Letter-Word Identification and Math Reasoning

subtests from the Woodcock-Johnson Psychoeducational Battery—Revised (WJ-R; Woodcock & Johnson, 1989). Standard scores based on the WJ-R normative sample were utilized to estimate the child's academic achievement in comparison to other children the same age. In addition, underachievement in reading or math relative to intelligence was assessed by subtracting *z*-scores for mathematics reasoning or word identification from *z*-scores for intelligence, controlling for regression effects using the formula provided by Frick *et al.* (1991).

Teacher Report Measures

Each child's primary classroom teacher completed several age-appropriate measures of social competence. The preschool version of the Social Skills Rating System (SSRS; Gresham & Elliott, 1990) includes scales assessing Cooperation, Assertion, and Self-Control. The internal reliability of the SRSS scales is high ($\alpha = .93–.94$), and SSRS ratings are significantly associated with other measures of social competence and adaptive behavior (e.g., Flanagan, Alfonso, Primavera, Povall, & Higgins, 1996; Gresham & Elliot, 1990; Merrell, 1995).

The Teacher Assessment of Social Behavior (TASB; Cassidy & Asher, 1992) is a 12-item scale designed to assess dimensions of social functioning. This scale provides scores on four dimensions of social behavior (prosocial, shy/withdrawn, disruptive, and aggressive). The internal reliability of the four scales is adequate to high ($\alpha = .62$ for shy/withdrawn, .88 or higher for the other three scales). In a sample of kindergarten and first-grade children, children classified as low-accepted on the basis of peer sociometric ratings were found to have significantly lower scores on the prosocial behavior dimension and significantly higher teacher ratings of shy/withdrawn, disruptive, and aggressive social behavior.

Finally, teachers were asked to estimate the proportion of each child's peers who like, dislike, or ignore the child, using the procedure developed by Dishion (1990) and adapted for the DSM-IV field trials (Lahey *et al.*, 1994). Support for the validity of these ratings in this sample is provided by significant correlations with children's self-report of friendship difficulties ($r = .22–.35$) and teacher ratings on the SSRS ($r = .51–.59$) and TASB ($r = .52–.66$).

Child Report Measure

Each child completed a well-validated self-report instrument designed to assess the child's ability to make and keep friends, as well as the extent that the child feels left out of peer activities (Cassidy & Asher, 1992). Cassidy and Asher (1992) found that all but 1 of the 15 items from this measure loaded on a single factor, and that the internal reliability of the measure was satisfactory (Cronbach's $\alpha = .79$). Children with low ratings of social acceptance by peers were found to report significantly lower friendship scores than average or highly accepted children. Moreover, children with the lowest friendship scores were rated as significantly less prosocial and more aggressive by teachers and peers and as shier by peers.

CGAS

The parent and the interviewer each completed the nonclinician version of the CGAS, a rating scale of global adaptive functioning that ranges from 1 to 100 (Setterberg, Bird, & Gould, 1992). At each decile, the rater is provided with a phrase that describes functioning at that level (e.g., 1–10 = "extremely impaired; so impaired that constant supervision is required for safety"; 91–100 = "doing very well in all areas; no problems at home, at school, or with friends; likeable, confident, involved in activities. Functioning is superior or above average"). Raters are asked to provide the single number that best represents the child's lowest level of functioning during the past 6 months.

RESULTS

Internal Validity and Interrater Consistency on the HBRS

Because the HBRS has not been utilized to rate behavior during testing in any previous published studies, the psychometric qualities of the scale were assessed in some detail. Zero-order correlations among the three items comprising the AD/HD composite were high for both tester and interviewer ratings, $r(252) = .74–.81$, all $p < .001$, as were correlations between each item and the AD/HD composite score, $r(252) = .88–.93$, all $p < .001$. The internal consistency of the AD/HD composite was excellent for both the tester ($\alpha = .93$) and the interviewer ($\alpha = .92$) ratings.

Interrater consistency was assessed by examining the correlations between the ratings of the tester and the interviewer. Correlations between raters ranged from moderate to high for the three items, $r(252) = .58–.68$, all $p < .001$, and the correlation between the AD/HD composite scores of the interviewer and the tester was .76($p < .001$). These significant correlations suggest that the HBRS has adequate interrater reliability.

Table II. Correlations Between Hillside Behavior Rating Scale Ratings and Parent and Teacher Reports[a] of Attention-Deficit/Hyperactivity Disorder Symptoms

	HBRS tester rating		HBRS interviewer rating	
Measure of DSM-IV AD/HD	Zero-order correlation	Partial correlation[b]	Zero-order correlation	Partial correlation[b]
Parent report				
Hyp/Imp symptoms	.50[c]	.35[c]	.51[c]	.35[c]
Inattention symptoms	.46[c]	.26[c]	.46[c]	.27[c]
Total AD/HD symptoms	.50[c]	.34[c]	.50[c]	.34[c]
Teacher report				
Hyp/Imp symptoms	.37[c]	.22[c]	.40[c]	.24[c]
Inattention symptoms	.32[c]	.19[d]	.31[c]	.18[d]
Total AD/HD symptoms	.38[c]	.19[d]	.39[c]	.21[d]

[a] Parent report $n = 252$; teacher report $n = 247$.
[b] Partial correlation controlling intelligence, age, gender, and internalizing and externalizing symptoms.
[c] $p < .001$.
[d] $p < .01$, one-tailed.

Convergent Validity of the HBRS

The convergent validity of the HBRS was examined by testing if HBRS ratings were significantly associated with the number of symptoms of AD/HD reported by parents or teachers. The mean HBRS AD/HD score of participants with DSM-IV AD/HD (Interviewer $M = 6.91$, $SD = 3.27$; Tester $M = 6.79$, $SD = 3.09$) was significantly higher than the mean of the comparison sample (Interviewer $M = 4.29$, $SD = 1.95$; Tester $M = 4.27$, $SD = 1.90$), Interviewer $t(250) = 7.17$, Tester $t(250) = 7.04$, both $p < .001$. Similarly, zero-order correlations were significant between HBRS ratings and symptoms of inattention and hyperactivity/impulsivity reported by the child's mother or primary classroom teacher (Table II). HBRS scores also correlated significantly with symptoms of ODD as reported by parents, $r(252) = .34$, $p < .001$), and teachers, $r(247) = .24$, $p < .001$, and with symptoms of CD as reported by parents, $r(252) = .22$, $p < .001$.

To test if HBRS ratings were independently associated with AD/HD, partial correlations were calculated between HBRS scores and symptoms of each of the disruptive behavior disorders (AD/HD, ODD, and CD) while controlling for symptoms of the other two disorders, as well as age, gender, intelligence, and internalizing symptoms. Partial correlations between HBRS ratings and parent and teacher ratings of AD/HD symptoms remained significant when controlling for symptoms of ODD, CD, and the other covariates ($pr = .18$–$.33$, all $p < .01$). In contrast, nonsignificant partial correlations were obtained between HBRS ratings and all measures of ODD or CD when symptoms of AD/HD were controlled ($pr = .01$–$.04$, all $p = ns$), suggesting that HBRS ratings are independently associated with symptoms of AD/HD as reported by parents and teachers.

Concurrent Validity of the HBRS

A diagnostic measure can be said to have concurrent validity if higher scores are significantly associated with difficulties in other areas that are not assessed by the measure. The concurrent validity of the HBRS was tested by examining the relation between HBRS scores and the measures of functional impairment.

Zero-order correlations between HBRS scores and the impairment measures indicated that HBRS ratings were significantly associated with impairment in all domains with the exception of reading achievement (Table III). Partial correlations controlling age, gender, intelligence, and internalizing and externalizing symptoms remained significant for nearly all measures, although only tester ratings were significantly associated with the SSRS Cooperation scale, and only interviewer ratings were significantly associated with the SSRS Assertion scale.

Incremental Validity of HBRS Ratings

A series of hierarchical multiple regression analyses were conducted to test if HBRS ratings provided a significant increment in the prediction of functional impairment beyond the variance associated with parent and teacher reports. Intelligence, age, gender, and symptoms of other psychopathology were entered in the first step

Table III. Zero-Order and Partial Correlations Between HBRS Scores and Measures of Functional Impairment[a]

Measure of impairment	HBRS tester rating		HBRS interviewer rating	
	Zero-order correlation	Partial correlation[b]	Zero-order correlation	Partial correlation[b]
Cognitive measures				
Prorated Intelligence	−.28[d]	−.21[c,e]	−.28[d]	−.20[c,e]
WJ-R Math Reasoning	−.30[d]	−.13[f]	−.29[d]	−.13[f]
Intell./Math Discrepancy	.12[f]	.12[c,f]	.11[f]	.11[c,f]
WJ-R Letter Word ID	.08	−.01	.03	−.06
Intelligence/Reading Discrepancy	.03	.02[c]	.08	.07[c]
Impairment reported by teacher				
Peers like	−.37[d]	−.21[e]	−.34[d]	−.18[e]
Peers dislike	.31[d]	.14[f]	.19[e]	−.02
Peers ignore	.27[d]	.17[e]	.26[d]	.17[e]
SSRS Cooperation	−.30[d]	−.13[f]	−.28[d]	−.10
SSRS Aggression	−.24[d]	−.09	−.28[d]	−.13[f]
SSRS Self-Control	−.41[d]	−.24[d]	−.40[d]	−.23[d]
TASB Aggression	.32[d]	.13[f]	.33[d]	.12[f]
TASB Disruption	.39[d]	.23[d]	.43[d]	.25[d]
TASB Prosocial	−.35[d]	−.21[d]	−.32[d]	−.17[e]
TASB Shy/Withdrawn	.08	.04	.12[f]	.08
Impairment reported by parent CGAS	−.38[e]	−.17[d]	−.39[d]	−.15[d]
Impairment reported by child				
Friendship difficulties	.32[d]	.20[d]	.35[d]	.25[d]
Impairment reported by interviewer CGAS	−.55[d]	−.39[d]	−.57[d]	−.39[d]

[a] For parent-report and child-report measures, $n = 252$; for teacher-report measures, $n = 247$. CGAS = Children's Global Assessment Scale, HBRS = Hillside Behavior Rating Scale, SSRS = Social Skills Rating System, TASB = Teacher Assessment of Social Behavior, and WJ-R = Woodstock-Johnson Psychoeducational Battery—Revised.

[b] Partial correlation controlling intelligence, age, gender, and internalizing and externalizing symptoms.

[c] Partial correlation controlling age, gender, and internalizing and externalizing symptoms.

[d] $p < .001$, one-tailed.

[e] $p < .01$, one-tailed.

[f] $p < .05$, one-tailed.

of the regression, to ensure that any significant findings were indicative of a specific relation between HBRS ratings and the measures of functional impairment. Parent and teacher reports of AD/HD symptoms were entered in the second step to test if these ratings were significantly associated with impairment. Finally, HBRS ratings by either the tester or the interviewer were entered in the third step to determine if HBRS ratings provided a significant increase in the explained variance of the impairment measures. Clinic ratings from the two examiners were entered in separate models so that the second set of regressions utilizing the interviewer ratings could serve as a replication of the initial models that included the tester ratings. In addition, it seemed probable that, in many clinical settings, two clinicians might not be available to rate a child's behavior, and so, these analyses allowed us to test separately the incremental validity of ratings by the tester and the interviewer.

Results of the regression analyses are summarized in Table IV. As expected, parent and teacher ratings accounted for a large proportion of the variance in the functional impairment measures. In addition, results of the separate multiple regression models indicated that ratings from both the tester and the interviewer provided a significant independent increment in the prediction of impairment in several domains. Specifically, clinic ratings provided unique information regarding children's prosocial behaviors, aggressive and disruptive behaviors, problems with self-control, and difficulties making and keeping friends. Note, however, that all significant findings for parent or teacher reports in Step 2 remained significant when HBRS ratings were added to the model. Moreover, parent and teacher reports account for a larger proportion of the variance than HBRS ratings for all measures except the child's report of friendship difficulties.

Table IV. Hierarchical Multiple Regression Models of the Independent Contribution of HBRS Ratings to the Prediction of Functional Impairment

	Step 1		Step 2			Step 3[a] HBRS Ratings			
	Covariates entered[b]		Parent	Teacher		Tester		Interviewer	
Measure of impairment	R^2	F	b	b	ΔR^2	b	ΔR^2	b	ΔR^2
Cognitive measure									
Prorated IQ[c]	.07	4.48[d]	−.37[e]	−.04	.077[e]	−.13	.012[f]	−.12	.010[g]
WJ-R Math Reasoning	.44	37.48[e]	−.07	−.08	.008	−.11	.008[f]	−.10	.007[f]
WJ-R Letter Word ID	.27	17.44[e]	−.15[f]	.08	.009	.09	.000	−.06	.002
Impairment reported by teacher									
Peers like	.27	17.71[e]	−.08	−.50[e]	.153[e]	−.10	.008[f]	−.06	.002
Peers dislike	.26	16.97[e]	−.07	.42[e]	.101[e]	.10	.007[g]	.09	.007[g]
Peers ignore	.10	5.38[e]	.16[f]	.36[e]	.124[e]	.07	.004[g]	.06	.002
SSRS Cooperation	.27	18.00[e]	−.14[f]	−.61[e]	.281[e]	.03	.000	.05	.004
SSRS Assertion	.16	10.03[e]	.23[d]	.23[d]	.079[e]	.01	.000	.03	.001
SSRS Self-Control	.33	24.66[e]	.03	−.48[e]	.143[e]	−.18	.023[d]	−.15	.016[d]
TASB Aggression	.33	39.24[e]	−.27[e]	.37[e]	.076[e]	.14	.012[d]	.12	.009[f]
TASB Disruption	.34	24.79[e]	−.06	.73[e]	.321[e]	.12	.010[d]	.15	.014[d]
TASB Prosocial	.30	20.69[e]	.01	−.46[e]	.140[e]	−.10	.007[f]	−.05	.002
TASB Shy/Withdrawn	.03	1.69	.20[f]	.01	.020[g]	−.02	.000	.03	.001
Impairment reported by parent CGAS	.46	42.01[e]	−.58[e]	.04	.150[e]	.01	.000	−.03	.000
Impairment reported by child									
Friendship difficulties	.11	7.27[e]	.08	.19[d]	.032[f]	.19	.026[d]	.23	.036[d]
Impairment reported by interviewer CGAS	.49	48.74[e]	−.54[e]	−.07[g]	.159[e]	−.18	.025[e]	−.23	.037[e]

Note. CGAS = Children's Global Assessment Scale, HBRS = Hillside Behavior Rating Scale, SSRS = Social Skills Rating System, TASB = Teacher Assessment of Social Behavior, and WJ-R = Woodstock-Johnson Psychoeducational Battery—Revised.

[a] Separate regressions were conducted using either the tester or interviewer rating as the third step.

[b] Covariates entered in the first step included IQ, age, gender, and symptoms of Oppositional Defiant Disorder, Conduct Disorder, MDD, and GAD.

[c] Covariates entered in the first step included age, gender, and internalizing and externalizing symptoms.

[d] $p < .01$.

[e] $p < .001$.

[f] $p < .05$.

[g] $p < .10$.

DISCUSSION

This study examined the validity of ratings of children's behavior during testing as part of a comprehensive AD/HD evaluation. The internal, convergent, concurrent, and incremental validity of ratings on the HBRS were examined in matched samples of preschool children with and without DSM-IV AD/HD. In the following sections we provide a brief summary of the current results, and relate these findings to results previously reported in the literature.

Internal and Convergent Validity of HBRS Ratings

The internal reliability of the HBRS was high and interrater reliability was adequate, suggesting that trained nonclinician examiners are capable of reliably rating symptoms of AD/HD. HBRS ratings were significantly associated with symptoms of DSM-IV AD/HD as reported by parents and teachers, providing support for the convergent validity of HBRS ratings. The significant association between HBRS ratings and teacher ratings of AD/HD is consistent with preliminary findings described by Barkley (1998) in a separate clinic-referred sample of children with AD/HD. In contrast, parent ratings were not significantly associated with ratings of clinic behavior in the sample described by Barkley, whereas HBRS scores were significantly related to parent ratings in the current sample. This discrepancy may reflect differences in the measures utilized to assess symptoms of AD/HD at home or in the clinic, or may be attributable to the fact that Barkley (1998) utilized a categorical approach for statistical analyses,

whereas the current report conducts stepwise multiple regressions using continuous data.

Clinical Utility of the HBRS

The finding that HBRS ratings of the child's behavior during testing are reliable and are associated with other measures of AD/HD provides preliminary support for the validity of the ratings, but furnishes no information regarding the clinical relevance of the observations. Ratings of behavior in the clinic setting have clinical utility only if these ratings significantly predict associated functional impairment, and provide at least some new information over and above the data provided by parents and teachers. In this sample, HBRS ratings were significantly associated with all measures of functional impairment with the exception of reading achievement, and provided a small but significant increment in the prediction of several impairment domains. Specifically, clinic observations were uniquely related to children's prosocial behaviors, problems with self-control, aggression, difficulties making and keeping friends, and academic achievement in mathematics, after controlling for all other variables. Associations between HBRS ratings and the measures of functional impairment were highly consistent across the two raters, providing a confirmation of findings and indicating that the significant relation with impairment was not restricted to the specific clinic settings in which each examiner observed the child.

In sum, this study suggests that ratings of AD/HD behaviors during testing are reliable, are significantly associated with other measures of AD/HD and measures of functional impairment, and provide an increment in the prediction of impairment over parent and teacher report alone. Therefore, HBRS ratings of behavior during testing may provide a source of unique information that may be helpful as one component of a comprehensive clinical evaluation. Reports from either parents or teachers, however, provide a significant increment in the prediction of impairment over the prediction provided by the combination of the other rater and the HBRS ratings. Therefore, we wish to emphasize that although clinic ratings may provide a useful supplement to a full assessment battery, these ratings should *not* be utilized as a replacement for reports from either parents or teachers.

Limitations and Directions for Future Research

1. In contrast to the Likert-scale format ("not at all" to "very much") of most AD/HD rating scales, the HBRS utilizes specific behavioral descriptions as the anchors for each item. Therefore, it is conceivable that the unique structure of the HBRS is essential to the incremental validity of the clinic ratings. Future studies should test if the current findings can be replicated when utilizing other rating scales. Moreover, the DSM-IV dimensions of hyperactivity, impulsivity, and inattention are each assessed by only one item on the HBRS. Future studies of the reliability, validity, and clinical utility of ratings of the specific DSM-IV symptoms during testing would provide a useful extension to the present results.
2. HBRS ratings provide a significant increment above parent and teacher ratings in the prediction of a variety of domains of functional impairment. This finding suggests that, in research applications, HBRS ratings may be useful as an additional manifest indicator of the latent construct of AD/HD. The size of the increment in prediction was relatively small for most impairment measures, however, suggesting that additional research needs to be conducted in larger samples to determine how to best utilize HBRS ratings in the clinical setting. Elevated HBRS ratings could represent a marker for a specific diagnostic subtype of AD/HD, or might identify cases with more severe AD/HD symptomatology or higher levels of comorbidity. Furthermore, if future studies provide sufficient normative data, HBRS scores or other ratings of clinic behavior could provide a useful third data point for cases in which parent and teacher reports are highly divergent, or cases in which the parent and teacher ratings fall in the borderline range for clinical diagnosis (i.e., five positive symptoms of hyperactivity/impulsivity or inattention).
3. Prior to rating the child on the HBRS, the interviewer administered the DISC interview to the mother. Therefore, although the interviewer was instructed to complete the HBRS based solely on the child's behavior, the ratings by the interviewer could be influenced by their knowledge of the mother's responses during the DISC. However, two factors suggest that knowledge of parent ratings did not significantly bias the HBRS ratings by the interviewer. Parent ratings of AD/HD symptoms correlated almost identically with HBRS ratings by the interviewer and the tester, suggesting that the relation between parent ratings and HBRS ratings was similar whether or not the rater knew the results of the DISC. In addition, because only one impairment measure was provided by the parent, it seems unlikely that the association between

HBRS ratings and the remaining measures of impairment is attributable to knowledge of parental ratings on the DISC.

4. The fact that the ratings of clinic behavior were completed by nonclinician examiners leaves open the possibility that the current findings may not generalize to clinicians. It seems likely, however, that the utilization of less-experienced raters provides a conservative test of the utility of ratings of clinic behavior, because more seasoned clinicians might be expected to be more proficient observers of relevant behaviors during testing. Future studies utilizing more experienced clinicians would enable this hypothesis to be tested directly.
5. In addition to behavioral ratings by examiners, behavioral observations by unobtrusive observers during testing represent another potentially fruitful area for future study. For example, Barkley, DuPaul, and McMurray (1990) found that behavioral observations coded from a videotape of a child's behavior during the administration of a continuous performance test (CPT) better discriminated between children with and without AD/HD than did the results of the CPT itself.
6. Like most samples of children with AD/HD ascertained through clinics (Hartung & Widiger, 1998), the participants in the current study were predominantly male. The small subset of female participants did not provide sufficient power to test definitively for gender differences in the utility of behavioral observations in the clinic setting. Therefore, future studies incorporating a larger number of females may wish to test if the reliability or validity of clinic ratings differs between males and females.
7. It seems probable that young children may exhibit behavior in the clinic setting that is more typical of their behavior at home and in school than older children. As children develop, they may become increasingly cognizant of the socially inappropriate nature of hyperactive or disruptive behaviors in the clinic setting, which may cause these behaviors to become less apparent among older individuals with AD/HD. Therefore, future studies of clinic behavior in older children would provide a useful extension of the present findings.

REFERENCES

Abikoff, H., & Gittelman, R. (1985). The normalizing effects of methylphenidate on the classroom behavior of ADDH children. *Journal of Abnormal Child Psychology, 13*, 33–44.

American Psychiatric Association (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., revised). Washington, DC: Author.

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

Barkley, R. A. (1988). Child behavior rating scales and checklists. In M. Rutter, H. Tuma, & I. Lann (Eds.), *Assessment and diagnosis in child psychopathology* (pp. 113–155). New York: Guilford Press.

Barkley, R. A. (1998). *Attention-deficit hyperactivity disorder: A handbook for diagnosis and treatment* (2nd ed.). New York: Guilford Press.

Barkley, R. A., DuPaul, G. J., & McMurray, M. B. (1990). Comprehensive evaluation of attention deficit disorder with and without hyperactivity as defined by research criteria. *Journal of Consulting and Clinical Psychology, 58*, 775–789.

Barkley, R. A., & Edelbrock, C. (1987). Assessing situational variation in children's behavior problems: The Home and School Situations Questionnaires. In R. Prinz (Ed.), *Advances in behavioral assessments of children and families* (Vol. 3, pp. 157–176). Greenwich, CT: JAI Press.

Cassidy, J., & Asher, S. R. (1992). Loneliness and peer relations in young children. *Child Development, 63*, 350–365.

Dishion, T. (1990). The peer context of troublesome child and adolescent behavior. In P. Leone (Ed.), *Understanding troubled and troubling youth*, Newbury Park, CA: Sage.

DuPaul, G. J. (1991). Parent and teacher ratings of ADHD symptoms: Psychometric properties in a community-based sample. *Journal of Clinical Child Psychology, 20*, 245–253.

DuPaul, G. J., & Barkley, R. A. (1992). Situational variability of attention problems: Psychometric properties of the Revised Home and School Situations Questionnaires. *Journal of Clinical Child Psychology, 21*, 178–188.

DuPaul, G. J., & Stoner, G. D. (1994). *ADHD in the schools*. New York: Guilford Press.

Flanagan, D. P., Alfonso, V. C., Primavera, L. H., Povall, L., & Higgins, D. (1996). Convergent validity of the BASC and SSRS: Implications for social skills assessment. *Psychology in the Schools, 33*, 13–23.

Frick, P. J., Kamphaus, R. W., Lahey, B. B., Loeber, R., Christ, M. A. G., Hart, E. L., & Tannenbaum, L. (1991). Academic underachievement and the disruptive behavior disorders. *Journal of Consulting and Clinical Psychology, 59*, 289–294.

Gittelman, R. G., & Klein, D. (1985). Hillside Behavior Rating Scale. *Psychopharmacology Bulletin, 21*, 898–899.

Gresham, F. M., & Elliott, S. N. (1990). *Social Skills Rating System: Preschool Level*. Circle Pines, MN: American Guidance Service.

Hart, E. L., Lahey, B. B., Loeber, R., & Hanson, K. S. (1994). Criterion validity of informants in the diagnosis of disruptive behavior disorders in children: A preliminary study. *Journal of Consulting and Clinical Psychology, 65*, 599–610.

Hartung, C. M., & Widiger, T. (1998). Gender differences in the diagnosis of mental disorders: Conclusions and controversies of the DSM-IV. *Psychological Bulletin, 123*, 260–278.

Hinshaw, S. P. (1994). *Attention deficits and hyperactivity in children*. New York: Sage.

Klein, R. G., & Abikoff, H. (1997). Behavior therapy and methylphenidate in the treatment of children with ADHD. *Journal of Attention Disorders, 2*, 89–114.

Lahey, B. B., Applegate, B., McBurnett, K., Biederman, J., Greenhill, L., Hynd, G. W., Barkley, R. A., Newcorn, J., Jensen, P., Richters, J., Garfinkel, B., Kerdyk, L., Frick, P. J., Ollendick, T., Perez, D., Hart, E. L., Waldman, I., & Shaffer, D. (1994). DSM-IV field trials for Attention Deficit Hyperactivity Disorder in children and adolescents. *American Journal of Psychiatry, 151*, 1673–1685.

Lahey, B. B., Pelham, W. E., Loney, J., Tripani, C., Stein, M., Lee, S., Kipp, H., Schmidt, E., Erhardt, A., Gold, E., Cale, M., Hartung, C., & Willcutt, E. (1999). *Comparison of five methods of using parent and teacher reports of symptoms in the diagnosis of DSM-IV attention-deficit/hyperactivity disorder*. Manuscript submitted for publication.

Lahey, B. B., Pelham, W. E., Stein, M. A., Loney, J., Trapani, C., Nugent, K., Kipp, H., Schmidt, E., Lee, S., Cale, M., Gold, E., Hartung, C. M., Willcutt, E., & Baumann, B. (1998). Validity of DSM-IV Attention-Deficit/Hyperactivity Disorder for young children. *Journal of the American Academy of Child and Adolescent Psychiatry, 37*, 695–702.

Merrell, K. W. (1995). Relationships among early childhood behavior rating scales: Convergent and discriminant construct validity of the Preschool and Kindergarten Behavior Scales. *Early Education and Development, 6*, 253–264.

Milich, R., Hartung, C. M., Martin, C. A., & Haigler, E. D. (1994). Behavioral disinhibition and underlying processes in adolescents with disruptive behavior disorders. In D. K. Routh (Ed.), *Disruptive behavior disorders in childhood* (pp. 109–138). New York: Plenum Press.

Molina, B. S. G., Pelham, W. E., Blumenthal, J., & Galiszewski, E. (1998). Agreement among teachers' behavior ratings of adolescents with a history of attention deficit hyperactivity disorder. *Journal of Clinical Child Psychology, 27*, 330–339.

Pelham, W. E., Gnagy, E. M., Greenslade, K. E., & Milich, R. (1992). Teacher ratings of DSM-III-R symptoms for the disruptive behavior disorders. *Journal of the American Academy of Child and Adolescent Psychiatry, 33*, 529–539.

Piacentini, J. C., Cohen, P., & Cohen, J. (1992). Combining discrepant information from multiple sources: Are complex algorithms better than simple ones? *Journal of Abnormal Child Psychology, 20*, 51–62.

Schwab-Stone, M. E., Shaffer, D., Dulcan, M. K., Jensen, P. S., Fisher, P., Bird, H. R., Goodman, S. H., Lahey, B. B., Lichtman, J. H., Canino, G., Rubio-Stipec, M., & Rae, D. S. (1996). Criterion validity of the NIMH Diagnostic Interview Schedule for Children Version 2.3 (DISC 2.3). *Journal of the American Academy of Child and Adolescent Psychiatry, 35*, 878–888.

Setterberg, S., Bird, H., & Gould, M. (1992). *Parent and interviewer version of the Children's Global Assessment Scale*. New York: Columbia University.

Shaffer, D., Fisher, P., Dulcan, M. K., Davies, M., Piacentini, J., Schwab-Stone, M. E., Lahey, B. B., Bourdon, K., Jensen, P. S., Bird, H. R., Canino, G., & Rieger, D. A. (1996). The NIMH Diagnostic Interview for Children Version 2.3 (DISC-2.3): Description, acceptability, prevalence rates, and performance in the MECA study. *Journal of the American Academy of Child and Adolescent Psychiatry, 35*, 865–877.

Shaffer, D., Fisher, P., Piacentini, J., Schwab-Stone, M., & Wicks, J. (1993). *Diagnostic Interview Schedule for Children*. New York: Columbia University.

Sleator, E. K., & Ullmann, R. K. (1981). Can the physician diagnose hyperactivity in the office? *Pediatrics, 67*, 13–17.

Thorndike, R., Hagan, E., & Sattler, J. (1986). *The Stanford-Binet Intelligence Scale* (4th ed.). Chicago: Riverside Press.

Woodcock, R. W., & Johnson, M. (1989). *Woodcock-Johnson Psychoeducational Battery—Revised*. Allen, TX: DLM Teaching Resources.