Using behavior rating scales for ADHD across ethnic groups: The IOWA Conners

By: Robert Reid, Charles D Casat, H James Norton, Arthur D Anastopoulos, and E Paige Temple

Reid, R., Casat, C.D., Norton, H. J., Anastopoulos, A.D., & Temple, E.P. (2001). Using behavior rating scales for ADHD across ethnic groups: The IOWA Conners. *Journal of Emotional and Behavioral Disorders, 9,* 210-218.

Abstract:

In this study we examined the normative and construct equivalence of the teacher IOWA Conners Rating Scale (IOWA) in a sample of 3,998 elementary school children (2,124 African American and 1,874 European American) ages 5 to 11 years in an urban school district. Risk odds ratios (% > 2 SD) were calculated by gender and ethnicity. An exploratory Principal Axis factor analysis was performed to determine the appropriateness of the 2-factor model. Structural equation modeling was used to estimate the degree of fit for the 2-factor model. Both African American boys and girls received significantly higher scores than their European American counterpoints.There was a 2.48 to 3.51 greater likelihood forAfrican American boys and a 3.60 to 5.27 greater likelihood of African American girls to be rated > 2 SD above the mean for inattention/overactivity, aggression, or IOWA Conners Rating Scale scores. A rater ethnicity by student ethnicity (European American vs. African American) interaction was also found. Confirmatory factor analysis indicated that the same 2-factor model was appropriate for the African American and European American groups. The results suggest that although there is construct equivalence across the African American and European American groups, there is still a question as to normative equivalence.

Article:

AN ESTIMATED 3% To 5% OF SCHOOLage children have attention-deficit/ hyperactivity disorder (ADHD). Children with ADHD are at a high risk for educational and behavioral problems (American Psychiatric Association, 1994). Almost 50% of children with ADHD will be placed in special education programs for learning disabilities and behavioral disorders (Reid, Maag, Vasa, & Wright, 1994). Aggression commonly co-occurs with ADHD. Aggressive behaviors, especially, have been demonstrated to have a high stability in childhood and adolescence (Campbell & Ewing, 1990; Loeber, 1990). Adverse outcomes among those individuals whose aggressive behaviors continue into later childhood and adolescence, including early school dropout, teenage pregnancy, delinquency, lowered occupational attainment, development of antisocial personality, substance abuse, and criminality in adulthood, have been well-documented (Loeber, 1990; Offord, 1989; Olweus, 1979). Thus, there is some urgency about identification and intervention, given the broad-ranging burden imposed in terms of individual and family suffering, lost educational opportunity, lower productivity, and economic impact on the family and the community as a whole.

Behavior rating scales are one of the most commonly used methods in the ADHD assessment process (Barkley, 1998; DuPaul & Stoner, 1994). Until the last decade, it was commonly believed that expression, course, and outcome of psychological disorders were largely universal and independent of cultural factors (Marsella & Kameoka, 1989). For this reason, the use of behavior rating scales across different ethnic groups has received scant attention in the past. If this premise is true, we would expect no differences across ethnic groups in prevalence rates or expression of ADHD. However, a growing body of literature suggests that cross-cultural differences may represent an important factor in assessment (Reid, 1995). Consequently, the applicability of behavior rating scales to culturally or ethnically diverse populations has increasingly been critically questioned (e.g., Baumeister, Berrios, Jiminez, Acevedos, & Gordon, 1990; Reid et al., 1998).

Recognition of cultural differences has thus become an important issue to be considered when undertaking programs of school-based risk identification that use screening instruments with minority children. The need for further research in this area is clear and compelling. It was estimated that by the year 2000, approximately one third of public school children would be from culturally different backgrounds (American Council on Education, 1988).Already, nearly half of the student population in our most populated cities and metropolitan areas are from culturally different groups (American Council on Education, 1988) and in two states, New Mexico and Mississippi, they constitute a majority (Quality Education for Minorities Project, 1990). With this increase in the number of culturally different children is a corresponding increase in the number of culturally different children with emotional or behavioral impairment. This, in turn, dictates a need to pay close attention to the assessment practices for culturally different children with special needs.

Based on a review of studies using ADHD behavior rating scales with culturally different groups, Reid (1995) concluded that (a) insufficient data existed to determine the extent to which psychometric properties of rating scales were consistent across different groups, (b) evidence suggested that some groups might be overidentified, (c) culturally different individuals were not adequately represented in the norm groups of many of the available scales, and (d) the possibility of rater bias could exist when individuals from one cultural group rate children from a different cultural group. He further cautioned that for culturally different individuals, normative comparisons may be misleading and the normative use of rating scales for identification of ADHD with culturally different individuals may be inappropriate. However, he also cautioned that the database was too small for any firm conclusions. When using assessment instruments with culturally different students, there are two major areas of concern:

1. Is there normative equivalence? Can the same norms be used for students from different ethnic groups? and
2. Is there construct equivalence? Do behavior ratings scales assess the same construct when used with different ethnic groups?

NORMATIVE AND CONSTRUCT EQUIVALENCE

Normative Equivalence

In terms of normative equivalence, there is now a well-documented pattern of significantly higher ratings for African American children as opposed to European American children (Reid et

al., 1998): Twice as many African American children screen positive for ADHD. Several possible interrelated explanations for the difference include rater effects, differences in socioeconomic status, real differences in behavior, halo effects, and a combination of all these factors.

Rater Effects. Rater effects refer to a situation in which ratings are systematically biased due to factors internal to the rater. Because behavior ratings reflect the subjective impressions of the rater (Barkley, 1987), rater effects can possibly occur in two ways: First, raters from different cultural groups may perceive behavior differently and thus differ in their ratings. Evidence supports differences based on rater ethnicity. Mann and colleagues (Mann et al., 1992; Mueller et al., 1995) asked mental health professionals and teachers to rate videotaped vignettes of children. They found that behavior ratings from Chinese and Indonesian mental health professionals and teachers were significantly higher than those of U.S. and Japanese mental health professionals and teachers. Because the actual behavior viewed remained constant, the results strongly suggest that the differences were due to the culture of the rater. The second possible cause of rater effects is when a rater from one cultural group rates a participant from a different cultural group. This situation occurs often. Because the majority of public school teachers are European American, it is common for teachers to rate children from a different cultural group. If a rater effect induces bias, which in turn results in spuriously high ratings related to the ethnicity of the child, it is a potentially serious problem.

The area of rater bias has received very little attention, and results are mixed. Some experimental evidence suggests the existence of biased ratings. Sonuga-Barke, Minocha, Taylor, & Sandberg (1993) assessed the extent to which teachers' ratings of behavior corresponded to actometer readings and behavioral observations in two experiments using Asian and English school-age children. The results of both experiments showed that although objectively measured behavior across the two groups was identical, teachers'ratings of Asian students were significantly higher than their English counterparts. In contrast, Jarvinen and Sprague (1995) used the ADD-H Comprehensive Teacher's Rating Scale (ACTeRS; Ullmann, Sleator, & Sprague, 1991) to assess whether items functioned differentially across different ethnic groups. They found that although some items were biased, no evidence indicated any systematic pattern of item bias that would inflate the scores of European American or African American groups.

Socioeconomic Status. Low socioeconomic status is a risk factor for one common behavior problem-ADHD (Biederman et al., 1995)-for several possible reasons. First, low socioeconomic status may be associated with other ADHD risk factors, such as severe marital discord, large family size, or foster care placement. Second, low socioeconomic status may expose children to environmental or psychosocial stressors. For example, Murphy and colleagues (1998) found that hunger resulted in impaired functioning and higher hyperactivity scores. These two factors might result in an actual increase in problem behaviors. Finally, socioeconomic status may affect observers' perceptions of behavior. Stevens (1981) used a simulation study in which teachers viewed identical videotapes of children. Along with the video segments, teachers were presented vignettes that described the children as either middle or low socioeconomic status. The low socioeconomic status description resulted in significantly higher hyperactivity ratings despite the fact that the actual behavior was identical.

Halo Effects. The presence of halo effect in scales assessing disruptive behavfor disorders has been well documented (e.g., Abikoff, Courtney, Pelham, & Koplewicz, 1993; Abikoff & Gittelman, 1985; Blunden, Spring, & Greenberg, 1974; Prinz, Connor, & Wilson, 1981; Schachar, Sandberg, & Rutter, 1986). This halo effect appears to be unidirectional in nature (Abikoff et al., 1993; Schachar et al., 1986). When children evidence oppositional or aggressive behaviors, raters tend to endorse items relating to hyperactivity or inattention, even when actual behaviors are not displayed. However, the opposite does not occur; hyperactivity or inattention do not result in inflated ratings of oppositional or aggressive behaviors. This results in spuriously inflated scores for hyperactivity and/or inattention. Therefore, if a given ethnic group actually displayed or was perceived to display aggressive or oppositional behaviors, then there is a distinct possibility that the result would be artificially inflated scores on unrelated areas (e.g., hyperactivity or inattention).

Construct Equivalence

Construct equivalence is a critical factor in cross-cultural assessment. If a given instrument functions differently (e.g., has a different factor structure) when used across different groups, then the scores across groups will not reflect the same construct and thus are not directly comparable. Little research is available on construct equivalence. Two studies have investigated the extent to which ADHD rating scales are equivalent (i.e., assess identical constructs) across African American and European American children. Reid et al. (1998) examined crosscultural equivalence of the 18-item ADHD Rating Scale-IV (School Version; DuPaul et al., 1997) for 381 African American boys and 1,359 European American boys. Results indicated a moderate degree of congruence across groups. There were an equal number of factors loading on similar items for both groups but differences in item, intercorrelations, and a disproportionately high percentage of African American boys screening positive for ADHD. Epstein, March, Conners, and Jackson (1998) used the Conners Teacher Rating Scale to examine cross-cultural equivalence for 609 European American and 418 African American students. They also found that there were similar factors across groups; however, groups differed in the presence of an antisocial factor for African American boys and an inattention factor in European American girls. Raters (i.e., classroom teachers) also tended to rate African American children higher on extemalizing behaviors.

Research in cross-cultural assessment has been limited in terms of the number of different scales used. The applicability of one commonly used scale, the IOWA Conners (Pelham, Milich, Murphy, & Murphy, 1989), in terms of normative and construct equivalence for the screening of elementary school-age African American children has not been studied. The IOWA Conners has two subscales-Inattention/ overactivity (10), and Aggression (WA). Internal reliability and test-retest reliability for the IOWA Conners are good. Loney and Milich (1982) reported alphas of .87 and .83 respectively for the 10 and WA subscales in a classroom sample, and test-retest stability was .87 and .85 for the two subscales. The WA scale potentially enhances its screening usefulness because problems with aggression are common among children with ADHD (Barkley, 1998). The IOWA Conners has demonstrated its usefulness as a tool with which to screen for risk of externalizing behavioral disturbance (Atkins, Pelham, & Licht, 1988; Casat, Norton, & Boyle-Whitesel, 1999; Loney & Milich, 1982; Pelham, Milich, Murphy, & Murphy, 1989). For example, Atkins et al. found that high IOWA Conners' scores predicted negative 3-year outcomes, while Casat et al. found significant correlations between high 10, WA, and IOWA

scores and externalizing diagnoses, Teacher Report Form (TRF) and Child Behavior Checklist (CBCL) externalizing scores, and Child and Adolescent Functional Assessment Scale (CAFAS) scores. In this study we investigated whether there are cultural effects on IOWA Conners' scores across European American and African American children. We addressed three specific questions:

1. Does the IOWA Conners demonstrate construct equivalence across the different groups?
2. Does the IOWA Conners demonstrate normative equivalence across groups? 3. Does rater ethnicity (European American or African American) affect ratings?

METHOD

Participants

Participants in this study consisted of 3,998 children (2,124 African American and 1,874 European American) ages 5 to 11 years, drawn from nine urban elementary schools in the southeast region of the United States. All schools were defined as being "high-risk schools" on the basis of (a) percentage of low socioeconomic status students, as estimated by the number of students with free/reduced cost lunch programs; (b) percentage of students below grade level as measured by end-ofgrade testing; and (c) the number of students with excessive school absences. Mean free/reduced lunch status across the nine participating schools was 38.8% (range 19.7%-75.4%). Each school was surveyed in the spring of the school year, either March or April, when the teachers were acquainted with their students. The mean survey participation across schools was 77.7% (range 57.4%-88.1%). One hundred seventy-eight general education teachers took part in the survey, of which 76.4% were European American women, 18% African American women, 4.5% European American men, and 1.1% African American men. From 200 to 300 students were included for each age level. The number of participants varied somewhat across age levels. However, the results of a chi-square test showed that participants were proportionally distributed across age levels and ethnicity X^sup2(6, N = 3,998) = 8.7, p = .18.

Procedures

An inservice was conducted with the teachers at each school to familiarize them with completion of the IOWA Conners.

An informed consent form was sent home in the book bag of each child for parent notification of the survey, as well as a cover letter from the school's principal endorsing the survey. All children whose parents consented were included in the study. The teachers completed a bubblesheet-scannable version of the IOWA Conners on each eligible child in his or her class.

Statistical Analysis

Descriptive statistics included frequency distributions for 10 and WA subscales for the European American and African American groups and means and standard deviations for each group by age and gender. To test for construct equivalence, an exploratory Principal Axis factor analysis using varimax rotation for the European American and African American groups was performed to determine if the two-factor model was appropriate for both groups. To test the fit of the twofactor model across the groups, separate confirmatory factor analyses (using structural equation modeling and the LISREL 8 program) were performed for boys and girls in the African American and European American groups using polychoric correlation and asymptotic

covariance matrices with Weighted Least Squares estimation. Finally, structural equation modeling was used to estimate how well the two-factor model fit across African American and European American boys and girls using covariance matrices and Generalized Least Squares estimation. To examine normative equivalence, we tested for differences in the rate of positive screens across African American and European American groups and for mean differences across ethnicity and gender. To test for possible differences in identification rates across the European American and African American groups, we computed odds ratios for boys and girls for the 10, WA, and IOWA in each group using the ratio of percentage African American to percentage European American. To test for mean differences, ANCOVAs were computed for both 10 and WA subscales using age as a covariate to control for the effects of maturation. We also investigated the effects of ethnicity by comparing African American and European American ratings.

RESULTS

Table I shows the mean scores and standard deviations for the 10 and WA subscales by age, gender, and ethnicity. For the African American group, the 10 subscale means (and SDs) for girls and boys, respectively, were 3.94 (4.26) and 6.71 (4.96). For the European American group, means (and SDs) for the 10 subscale for girls and boys were 2.02 (3.16) and 4.14 (4.59), respectively. For the WA subscale, means (and SDs) for the African American group for girls and boys, respectively, were 2.85 (4.24) and 4.44 (5.06). For the European American group, means (and SDs) for the WA subscale for girls and boys, respectively, were 0.95 (2.52) and 2.08 (3.69). As expected, frequency distributions for 10 and WA produced positive skewness and, in addition, were more highly kurtotic (i.e., more scores were in the tail) for the African American groups for both the 10 and WA scales.

Construct Equivalence

We first conducted an exploratory principal axis factor analysis using varimax rotation for the European American and African American groups to determine if the two-factor model was appropriate. Table 2 shows the results of the factor analysis. Results suggest that the twofactor model is generally appropriate for both groups. Interestingly, despite the fact that the subscales were constructed to be uncorrelated, some items loaded on both factors. Next, we used structural equation modeling (SEM; LISREL 8) to compare whether the same two-factor model structure was equivalent across both African American and European American groups. Our analyses followed procedures suggested by Joreskog and Sorbom (1993). We performed separate confirmatory factor analyses for boys and girls in African American and European American groups using polychoric correlations and asymptotic covariance matrices with Weighted Least Squares estimation. Examination of modification indices suggested that for two items on the 10 scale (Items 4 & 5), a separate error covariance estimate was necessary. Separate error covariance estimates for these items were used for all groups. This was likely because the items were highly similar. Table 3 shows the results of these analyses. For all tests, the Goodness of Fit (GFI), Adjusted Goodness of Fit, and Confirmatory Fit Index (CFI) were all above the .90 level, which is indicative of acceptable fit. The Root Mean Square Error of Approximation was also below the .08 level, indicative of acceptable fit. Finally, we compared the model fit across African American and European American boys and girls using covariance matrices and Generalized Least Squares estimation. For the African American and European American boys, the GFI was .90, the Normed Fit Index (NFI) was .99, and the CFI was .99. Thus, the results

suggest that the same model is appropriate for both groups. Similar results were found for the African American and European American females. The GFI was .90, the NFI was .99, and the CFI was .99, all of which are indicative of acceptable fit. Thus, there was construct equivalence across the African American and European American groups for both boys and girls.

Normative Equivalence

Odds ratios for the percentage of group members above 2.0 SD were calculated for 10, WA, and IOWA by gender and ethnicity (Table 4). There was a 2.48 to 3.51 greater likelihood for African American boys, and a 3.60 to 5.27 greater likelihood of African American girls to be identified as > 2 SD above the mean for 10, WA, or IOWA scores than their European American counterparts. Thus, the African American group screened positive at a much higher rate than the European American group.

TABLE 1
Means and Standard Deviations for IO and WA by Age, Gender, and Ethnicity

	Girls				Boys			
	African American[a]		European American[b]		African American[c]		European American[b]	
Age	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
				For IO				
5	2.80	(3.80)	2.15	(3.04)	6.18	(5.29)	3.60	(4.48)
6	4.12	(4.21)	1.86	(3.12)	6.07	(5.01)	3.95	(4.67)
7	3.90	(4.18)	2.38	(3.47)	6.40	(4.75)	4.48	(4.67)
8	3.78	(4.08)	2.13	(3.40)	7.02	(5.07)	3.56	(4.21)
9	4.38	(4.40)	2.08	(3.14)	7.09	(5.08)	4.27	(4.49)
10	4.37	(4.46)	1.62	(2.65)	7.04	(4.80)	4.35	(4.91)
11	3.63	(4.51)	1.73	(3.20)	6.83	(4.95)	4.66	(4.61)
				For WA				
5	1.95	(3.60)	1.14	(2.74)	2.92	(4.46)	1.72	(3.38)
6	2.38	(3.71)	0.98	(2.63)	3.67	(4.75)	2.17	(4.01)
7	2.29	(3.87)	0.95	(2.44)	3.71	(4.42)	2.08	(3.68)
8	2.63	(4.03)	1.03	(2.57)	4.83	(5.53)	1.65	(3.44)
9	3.47	(4.68)	0.87	(2.48)	5.06	(5.08)	2.49	(3.76)
10	3.87	(4.79)	0.99	(2.71)	5.27	(5.29)	2.02	(3.69)
11	3.11	(4.35)	0.57	(1.77)	4.91	(5.30)	2.40	(3.71)

Note. IO = Inattention/Overactivity; WA = Aggression.
[a]n = 1,038. [b]n = 937. [c]n = 1,086.

TABLE 2
Item Loadings for African American and European American Groups

	European American		African American	
Item	**WA factor**	**IO factor**	**WA factor**	**IO factor**
IO1	.32	.81	.33	.80
IO2	.40	.64	.37	.66
IO3	.53	.63	.53	.65
IO4	.24	.88	.27	.85
IO5	.24	.77	.24	.78
WA1	.81	.37	.82	.36
WA2	.71	.31	.77	.28
WA3	.83	.27	.82	.31
WA4	.88	.26	.88	.31
WA5	.82	.34	.82	.36

Note. WA = Aggression; IO = Inattention/Overactivity.

TABLE 3
Results of Confirmatory Factor Analysis

	Fit indices			
Group	**GFI**	**AGFI**	**RMSEA**	**CFI**
AA Boys	.99	.99	.068	.99
AA Girls	.99	.99	.057	.99
EA Boys	1.00	.99	.059	1.00
EA Girls	.99	.99	.052	.99

Note. AA = African American; EA = European American; GFI = Goodness of Fit Index; AGFI = Adjusted Goodness of Fit Index; RMSEA = Root Mean Square Error of Approximation; CFI = Confirmatory Fit Index.

To test for differences across ethnicity and gender, we computed ethnicity (European American x African American) by gender (boy x girl) ANCOVAs for both the 10 and WA subscales, using age as a covariate. Before conducting ANOVAs, the covariate (age) was tested for homogeneity of regression. In the case of the 10 scale, there were no significant two- or three-way interactions with factors. Results showed that age was significantly related to 10 total score, $F(1, 3993) = 13.33$, $p < .001$. There were significant main effects for ethnicity, $F(1, 3993) = 273.49$, $p < .001$, with the scores for the African American group higher than the European American group, and for gender, $F(l, 3993) = 320.83$,$p < .001$, with scores higher for boys than for girls. Effect sizes for ethnicity (eta2 = .063) and gender (eta2 = .073) were moderate. In addition, there was a significant gender by ethnicity interaction, $F(1, 3993) = 5.63$,$p = .018$. However, the eta^sup2 value for the interaction was only .001, suggesting it is of little practical significance and is most likely due to the high power.

For the WA scale, there was a significant age-by-ethnicity interaction, $F(1, 3993) = 5.63$, $p < .001$). Because the assumption of homogeniety of regression was not met for one factor, we report on main effects for gender and consider ethnicity in terms of the age x ethnicity interaction. For gender, a significant main effect, $F(l, 3993) = 112.79$, $p < .001$, was found. The effect size was small to moderate (eta 2 =.027). There was a significant effect for ethnicity, $F(1, 3984) = 241.82$, $p < .001$, with a moderate effect size (eta^sup^2 = .057). The interaction between ethnicity and age was due to a disproportional increase in WA scores for the African American group as age increased. The effect size was small (eta2 = .007). Analysis of the increase in means in Table I show that there is a different pattern across gender and ethnicity. There is an increase of nearly 2 points for the African American boys ages 5 to 11, while the European American boys' increase was much lower. However, European American girls actually decreased slightly from ages 5 to 11, while their African American counterpoints increased. In the case of the African American boys, the magnitude of the increase suggests that it may have practical significance.

TABLE 4
Odds Ratios for IOWA, IO, and WA

	% AA (> 2SD)	% EA (> 2SD)	Odds ratio (AA/EA)
	Boys		
IOWA	12.98	5.24	2.48
IO	12.89	5.13	2.51
WA	14.64	4.17	3.51
	Girls		
IOWA	5.01	1.39	3.60
IO	3.95	0.75	5.27
WA	7.23	1.93	3.77

Note. IOWA = IOWA Conners Rating Scale (Pelham, Milich, Murphy, & Murphy, 1989); IO = Inattention/Overactivity; WA = Aggression; AA = African American; EA = European American.

To test for possible differences across raters, the ratings of women African American and European American teachers for the IO and WA scales using a 2 (teacher ethnicity) x 2 (student ethnicity) ANCOVA (with age as the covariate) were compared. Due to extremely small numbers, men teachers were excluded. Because effects of ethnicity were reported previously, we will report only effects of teacher ethnicity and interactions. For the IO scale there was no main effect for teacher ethnicity, F(l, 3774) = .64, p = .423. The interaction was significant, F(1, 3774) = 14.08, p < .001. The effect size was small (eta^sup2 = .004). A similar pattern was observed for the WA scale. Again, there was no main effect for teacher ethnicity, F(l, 3774) = 1.72, p = .190. The interaction was significant, F(1, 3774) = 10.67, p < .001, and again the effect size was small (eta^sup2 = .003). For both 10 and WA scales, the interaction was caused by a greater difference between the European American and African American groups for the European American teachers than for the African American teachers. To assess the potential impact of the interaction, we computed odds ratios for IOWA scores (% > 2SD for African American teachers/ % > 2SD for European American teachers) for all students by ethnicity and gender. For the African American boys and girls the odds ratios were .83 (11.16%/ 13.48%) and .58 (3.20%/5.49%), respecand girlS, odds ratios were I2.24(9.60-/ 4.41%) and 2.11 (2.45%/1.16%), respectively. The difference between teacher groups was most pronounced for the European American students and the African American girls.

DISCUSSION

The findings of this study of the IOWA Conners are consistent with those of previous research studies that have investigated the use of behavior rating scales across culturally different groups. Three main findings here are of interest: First, there appears to be construct equivalence across European American and African American groups. Second, normative equivalence is questionable; differences exist in distributions of IOWA scores and mean differences across the African American and European American groups, which leads to an increased likelihood for a positive screen for children in the African American group. Third, there were statistically significant rater ethnicity by student ethnicity interactions for the IO and WA subscales.

Construct Equivalence

Significant differences in factor structures or item loadings would make normative use across different groups problematic because comparisons would be based on different constructs. This is an important prerequisite for use of behavior rating scales across different cultural groups. In this study, exploratory factor analyses and confirmatory factor analyses substantiated the basic two-factor model for both groups. Thus the IOWA appears to measure the same construct across both European American and African American groups. This is consistent with previous research (Epstein et al., 1998; Reid et al., 1998) that used different scales (i.e., the Conners and the ADHD Rating Scale-IV) and further supports the idea that there is construct equivalence across European American and African American groups. The similar pattern of results across different scales provides convergent evidence that strongly suggests that construct equivalence is not a problem that would affect the use of behavior rating scales.

Normative Equivalence

Differences in mean scores and distributions across African American and European American groups have been previously reported for the Conners and the ADHD RS-IV (Epstein et al., 1998; Reid, 1995; Reid et al., 1998). This study adds the IOWA to the list of behavior rating scales that have documented significantly higher scores for African American children as opposed to European American children and strongly suggests that previously reported differences are not scalespecific. Consistent with previous studies (e.g., Reid et al., 1998), the effect size of differences across African American and European American groups was in the moderate range. Additionally, a similar pattern of distributional differences with greater kurtosis in the African American group was also observed. This resulted in a greatly increased likelihood for children in the African American group to screen positive. On average, African American boys were approximately 2.5 times more likely to screen positive, while African American girls were more than 3.5 times more likely to screen positive. Thus, there is a marked imbalance across the groups for both genders.

Due to limitations in the study, we cannot determine exactly which factor or combination of factors (e.g., rater effects, differences in socioeconomic status, real differences in behavior, halo effects) could account for the differences found. However, the results do allow for some inferences on the effects of two of these factors. First, all participants were selected from schools with a high proportion of atrisk, low socioeconomic status children. Unless a systematic difference in participation existed across the European American and African American groups (where only higher socioeconomic European American children and lower socioeconomic African American students participated), there should not be pronounced difference across groups on this factor. Moreover, other studies that have compared Hispanic students to European

American students have found minimal or no difference (e.g., DuPaul et al., 1997; Jarvinen & Sprague, 1995). If socioeconomic status alone accounted for differences, then we would expect to see the same pattern for Hispanic groups. This is not the case. Thus, it appears that socioeconomic status alone could not account for the observed differences. Second, the combination of well-documented halo effects for children with aggressive or oppositional behavior and the higher scores on the WA scale for the African American group suggest the possibility, if not the likelihood, that inflated 10 scores are due to halo effects. The ethnicity by age interaction is also a source of concern. The disproportional increase in WA scores for the African American group as age increased suggests that the African American children are exhibiting (or are perceived as exhibiting) increased aggression with age. As a result, the African American group will be more likely to screen positive on the WA scale as age increases.

We would caution that in this study, as with many other studies investigating cultural differences, the lack of observational data and individual socioeconomic status data present distinct limitations in the interpretation of results. Because there are no empirical data on the actual behaviors exhibited by participants, we cannot exclude the possibility that the African American group actually exhibited higher rates of inattentive, hyperactive, and/or aggressive behaviors. Equally, we were not able to examine for the presence or absence of halo effects or the effects of socioeconomic status and its effects on other variables. However, either of these scenarios are cause for concern. If the ratings do reflect an actual difference in behaviors, then African American groups are at higher risk for behavior problems and should be targeted for intervention and increased treatment resources. If, on the other hand, some or much of the differences are due to halo effects, rater effects, or socioeconomic status, then there is a need for separate norms for African American students. We would suggest that the results of this study, when combined with results from previous studies, indicate that there is reason for caution when considering the screening application of the IOWA with African American children for purposes of risk identification and intervention planning.

Rater Effects

Although there were no main effects for teacher ethnicity for either the IO or WA subscales, there were significant teacher ethnicity by student ethnicity interactions. For both the IO and WA subscales, teachers rated the African American students higher than European American students; however, African American teachers on average rated African American students somewhat lower on both scales. Thus, it would appear that African American teachers tend to perceive less difference between African American and European American students than do European American teachers. On average, African, American children would more likely be rated higher if they were rated by a European American teacher as opposed to an African American teacher. Conversely, European American children would likely be rated lower if rated by a European American teacher than if rated by an African American teacher. This is most pronounced in the case of the WA subscale. It is also possible that the increased WA score, which reflects perceived aggression, could inflate the 10 scores due to halo effects.

Interpretation of these results is not straightforward. The results are consistent with the very limited body of research on rater effects (Mann et al., 1992; Mueller et al., 1995; Sonuga-Barke et al., 1993). Yet, it is possible that the interactions are simply an anomaly due to the high power. This seems unlikely due to the fact that the interaction occurred for both scales and was similar

(i.e., lower ratings for the African American teachers) for both scales, but we cannot rule it out. It is also possible that no practical significance exists. The differences in ratings were not large, and the effect sizes were small. However, effect sizes are largely concerned with children's mean scores, and the children of greatest concern are those who lie in one tail, that is, students who would screen positive. There were disparities across African American and European American groups in terms of positive screens (i.e., % > 2SD). The difference across raters was not pronounced for the African American boys; however, there was a 2% difference between African American and European American teachers. Interestingly, the most pronounced disparity across raters was with African American girls and European American students. African American girls were much more likely to screen positive when rated by European American teachers. European American students were much less likely to screen positive when rated by European American teachers. Taken as a whole, the data suggest the possibility of rater effects. However, the lack of behavioral data precludes any firm conclusions. These results do suggest that there is a pressing need to conduct studies akin to Sonuga-Barke et al. (1993) that combine behavior ratings and actual observed behavior.

Implications for Practice

Policymakers and researchers have advocated that mental health interventions would be more readily accessible and achieve greater effectiveness if greater resources were directed into school-based mental health services. (Jensen, Hoagwood, & Petti, 1996; Leaf et al., 1996). Implementation of prevention and early intervention services in school settings would in turn be aided by the use of screening instruments that allow for accurate, rapid identification of children who have significant behavioral problems or who are at high risk for early development of such difficulties. However, inherent in use of this school-based screening method is the need for examination of the appropriateness of the instruments themselves and their potential for introducing bias. This is especially pertinent for our urban school systems, with their large minority populations and heightened need for an array of special services, if the difficulties of children are not to be multiplied unwittingly. The current study with the IOWA Conners of the pattern of higher rates of positive screen identification of African American children, as with other instruments studied previously, serves to emphasize that caution is indicated in the application for purposes of large-scale risk screening. The potential for rater effects only accentuates the need for caution. Fundamental to the utility of behavior rating scales is the expectation that they have high predictive accuracy, to avoid both false positive and false negative labeling. The results of this study suggest that there is the possibility of false positives based on the combination of rater and student ethnicity. Further studies are required to elucidate the sources and explanations of these differences before behavioral screening instruments may be used normatively with confidence across ethnic groups in the school setting.

References

Abikoff, H., Courtney, M., Pelham, W. E., & Koplewicz, H. S. (1993). Teachers' ratings of disruptive behaviors: The influence of halo effects. Journal of Abnormal Child Psychology, 21, 519-533.

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

American Council on Education and Education Commission of the States. (1988). One-third of a

nation: A report by the Commission on Minority Participation in Edu

cation and American Life. Washington DC: Author.
American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders. (4th ed.). Washington, DC: Author.
Atkins, M. S., Pelham, W. F., Jr., & Licht, M. (1988). The development and validation of objective classroom measure for the assessment of conduct and attention deficit disorders. Advancement of Behavioral Assessment of the Child and Family, 4, 3-31.
Barkley, R. A. (1987). The assessment of attention deficit-hyperactivity disorder. Behavioral Assessment, 9, 207-233.
Barkley, R. A. (1998). Attention deficit hyperactivity disorder: A handbook for diagnosis and treatment (2nd ed). New York: Guilford Press.

Baumeister, J. J., Berrios, V., Jiminez, A. L., Acevedos, L., & Gordon, M. (1990). Some issues and instruments for the assessment of attention-deficit hyperactivity in Puerto Rican children. Journal of Clinical Child Psychology, 19, 9-16.
Biederman, J., Milberger, S., Faraone, S. V, Kiely, K., Guite, J., Mick, E., et al. (1995). Family-environment risk factors for attention-deficit hyperactivity disorder. Archives of General Psychiatry, 52, 464 64 470.
Blunden, D., Spring, C., & Greenberg, I. M. (1974). Validation of the Classroom Behavior Inventory. Journal of Consulting and Clinical Psychology, 42, 84-88.
Campbell, S. B., & Ewing, L. J. (1990). Follow-up of hard-to-manage preschoolers: Adjustment at age 9 and predictors of con

tinuing symptoms. Journal of Child Psychology and Psychiatry, 31, 871-889.
Carat, C. D., Norton, H. J., & Boyle-Whitesel, M. (1999). Identification of elementary school children at risk for disruptive behavioral disturbance: Validation of a combined screening method. Journal of the American Academy of Child and Adolescent Psychiatry, 38, 1246-1253.
DuPaul, G. J., Power, T. J., Anastopoulos, A. D., Reid, R., McGoey, K. E., & Ikeda, M. J. (1997). Teacher ratings of attention deficit hyperactivity disorder symptoms: Factor structure and normative data. Psychological Assessment, 9, 436-444.
DuPaul, G. J., & Stoner, G. (1994). ADHD in the schools: Assessment and intervention strategies. New York: Guilford.
Epstein, J. N., March, J. S., Corners, C. K., & Jackson, D. L. (1998). Racial differences on the Corners Teacher Rating Scale. Journal of Abnormal Child Psychology, 26, 109118.

Jarvinen, D. W., & Sprague, R. L. (1995). Using ACTeRS to screen minority children for ADHD: An examination of item bias. Journal of Psychoeducational Assessment, Special Issue on ADHD, 13, 172-184.
Jensen, P S., Hoagwood, K., & Petti, T. (1996). Outcomes of mental health care for children and adolescents II: Literature review and application of a comprehensive model. Journal of the American Academy of Child and Adolescent Psychiatry, 35, 1064-1077.
Joreskog, K., & Sorbom, D. (1993). LISREL 8. Hillsdale, NJ: Erlbaum.

Leaf, P. J., Algeria, M., Cohen, P., Goodman, S. H., Horwitz, S. M., Hoven, C. W., et al. (1996). Mental health service use in the community and schools: Results from the four-community

MECA study. Journal of the American Academy of Child and Adolescent Psychiatry, 35, 889-897.
Loeber, R. (1990). Development and risk factors of juvenile antisocial behavior and delinquency. Clinical Psychology Review, 10, 1-41.

Loney, J., & Milich, R. (1982). Hyperactivity, inattention and aggression in clinical practice. In M. Woolraich (Ed.), Advances in developmental and behavioral pediatrics. (Vol. 3, pp. 113-147). Greenwich, CT: JAI.
Mann, E. M., Ikeda, Y., Mueller, C. W., Takahashi, A., Tao, K. T., Humris, E., et al. (1992). Cross-cultural differences in rating hyperactive-disruptive behaviors in children. American Journal of Psychiatry, 149, 1539-1542.
Marsella, A. J., & Kameoka, V. A. (1989). Ethnocultural issues in the assessment of

psychopathology. In S. Wetzler (Ed.), Measuring mental illness in psychometric assessment for clinician (pp. 231-256). Washington, DC: American Psychiatric Press.
Mueller, C. W., Mann, E. M., Thanapum, S., Humris, E., Ikeda, Y., Takahashi, A., et al. (1995). Teachers' ratings of disruptive behavior in five countries. Journal of Clinical Child Psychology, 24, 434-442.
Murphy, J. M., Wehler, C., Pagano, M. A., Little, M., Kleinman, R. E., & Jellinek, M. S. (1998). Relationship between hunger and psychosocial functioning in low-income American children. Journal of theAmerican Academy of Child and Adolescent Psychiatry, 37, 163-170.
Offord, D. R. (1989). Conduct disorder: Risk

factors and prevention. In D. Shaffer, L. Phillips, & N. Enzer (Eds.), Manual of mental health, alcohol and other drug use in children and adolescents (pp. 273307; DHHS Publication ADM 89-1646), Washington, DC: Alcohol, Drug Use and Mental Health Administration.
Olweus, D. (1979). Stability of aggressive behavior patterns in males: A review. Psychological Bulletin, 86, 852-875.
Pelham, W. E., Jr., Milich, R., Murphy, D., & Murphy, H. A. (1989). Normative data on the IOWA Conners Teacher Rating Scale. Journal of Clinical Child Psychology, 18, 259-262.

Prinz, R. J., Connor, P. A., & Wilson, C. C. (1981). Hyperactive and aggressive behaviors in childhood. Intertwined dimensions. Journal of Abnormal Child Psychology, 9, 191-202.

Quality Education for Minorities Project. (1990). Education that works: An action plan for education of minorities. Cambridge: Massachusetts Institute of Technology.
Reid, R. (1995). Assessment of ADHD with culturally different groups: The use of behavioral rating scales. School Psychology Review, 24, 537-560.
Reid, R., DuPaul, G. J., Power, T. J., Anastopoulos, A. D., Rogers-Adkinson, D., Noll, M.-B., et al. (1998). Assessing culturally different students for attention deficit hyperactivity disorder using behavior rating scales. Journal of Abnormal Child Psychology, 26, 187-198.
Reid, R., Maag, J. W., Vasa, S. F., & Wright, G. (1994). Who are the children with ADHD: A school-based survey. Journal of Special Education, 28, 117-137.
Schachar, R., Sandberg, S., & Rutter, M. (1986). Agreement between teachers' ratings and observations of hyperactivity, inattentiveness, and defiance. Journal of Abnormal Child Psychology, 14, 331-345.
Sonuga-Barke, E., Minocha, K., Taylor, E., & Sandberg, S. (1993). Inter-ethnic bias in teachers'

ratings of childhood hyperactivity. British Journal of Developmental Psychology, 11, 187-200.
Stevens, G. (1981). Bias in the attribution of hyperkinetic behavior as a function of ethnic identification and socioeconomic status. Psychology in the Schools, 18, 99-106.
Ullmann, R. K., Sleator, E. K., & Sprague, R. L. (1991). ADD-H comprehensive teachers' rating scale-ACTeRS. Champaign, IL: MetriTech.