Assessing Culturally Different Students for Attention Deficit Hyperactivity Disorder Using Behavior Rating Scales

By: Robert Reid, George J. DuPaul, Thomas J. Power, Arthur D. Anastopoulos, Diana Rogers-Adkinson, Mary-Beth Noll, and Cynthia Riccio

Reid, R., DuPaul, G.J., Power, T.J., Anastopoulos, A.D., Rogers-Atkinson, D., Noll, M.B., & Riccio, C. (1998). Assessing culturally different students for AD/HD using behavior rating scales. *Journal of Abnormal Child Psychology*, *26*, 187-198.

Abstract:
Behavior rating scales are commonly used in the assessment of attention deficit-hyperactivity disorder (ADHD). However, there is little information available concerning the extent to which scales are valid with culturally different students. This study explored the use of the ADHD-IV Rating Scale School Version with male Caucasian (CA) and African American (AA) students from ages 5 to 18 years. Teachers rated AA students higher on all symptoms across all age groups. LISREL analysis indicated that scale does not perform identically across groups. This was supported by the results of multidimensional scaling with suggested that there is a different relation between items across groups. Implications for research and practice are discussed.
KEY WORDS: Attention deficit hyperactivity disorder; cross-cultural assessment; behavior rating scales.

Article:
In the United States, there is a pattern of disproportionate diagnosis and placement of African American, Hispanic, and Asian children in categories of disability. This pattern, first noted by Dunn (1968) and Mercer (1973), has persisted despite legal safeguards. For example, Chinn and Hughes (1987) analyzed data from 1978, 1980, 1982, and 1984 surveys performed by the Office of Civil Rights and found that African Americans were placed in classes for mild mental retardation at approximately twice the rate that would be expected and Hispanic children were underrepresented in categories of mild mental retardation, behavior disorders, and speech-language impairments. Recent data suggest that this trend is continuing (Hunt & Marshall, 1994). There are at least two reasonable explanations for disproportionate representation of some groups. One explanation is that culturally different individuals are more likely to be exposed to prenatal risk factors, psychosocial stressors, and economic disadvantage which in turn affect educational and behavioral outcomes. Alternatively, commonly used assessment instruments could be misleading or invalid when used with culturally different students.

Assessment of culturally different students with special needs has presented an ongoing problem. Students have been assessed using tests that were not presented in their native language *(Diana*

v. State Board of Education, 1970) which were biased (Sattler, 1988) or considered discriminatory *(Larry R v Riles*, 1979). There is a pressing need to address issues of cross-cultural assessment. At present, approximately 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).

One area that is drawing increasing attention is the assessment of attention deficit hyperactivity disorder (ADHD) with culturally different students (Reid, 1995). ADHD is now one of the most commonly diagnosed conditions of childhood (e.g., Epstein, Shaywitz, Shaywitz, & Woolston, 1991; Whalen, 1989) and accounts for approximately half of all referrals to child mental health clinics (Lerner & Lerner, 1991). Present estimates are that between 3 and 5% of school age children may manifest the symptoms of ADHD (American Psychiatric Association [APA], 1994; Barkley, 1990). Serious concerns have been expressed regarding the assessment of ADHD with culturally different students. Bauermeister, Berrios, Jimenez, Acevedos, and Gordon (1990) noted that both ADHD as a disorder and the instruments designed to assess it were derived from the perspective of Western professionals, using Western concepts of disorder and measurements, and without regard to cultural difference. This concern was echoed by the NAACP Legal Defense and Education Fund during the reauthorization hearing for PL 94142 (U.S. Congress, 1990). They expressed concerns that ADHD as a category of disability would "invite abuse for black children, especially black males, resulting in the disproportionate referral to special education" (Penning, 1990, p. 32). Therefore, because of (a) the increasing numbers of minority students, (b) the history or problems with assessments of culturally different groups, and (c) current concerns over the possibility of disproportionate diagnosis, there is a need to assess the confidence that we may have when assessing cultural minorities for ADHD.

One commonly used ADHD assessment instrument is the behavior rating scale (Barkley, 1990). One of the strengths of these instruments is that the large number of subjects used in developing rating scales allows for normative comparisons (Barkley, 1990). However, this does not guarantee that the norm group is representative of the population (Salvia & Ysseldyke, 1988). Instruments must also demonstrate reliability and validity for all the populations with which they are used. Because of the belief that etiology, expression, course, and outcome of psychological disorders were largely universal and independent of cultural factors (Marsella & Kameoka, 1989), cultural issues in the assessment of psychopathology have received scant attention. However, there is a growing body of literature that suggests that cross-cultural differences may be an important factor in assessment.

O'Donnell, Stein, Machabanski, and Cress (1982) used a modified Behavior Problem Checklist and demonstrated a lack of congruence between factor structures across Anglo-American and Mexican American groups. Weisz et al. (1987, 1989) reported significant differences in reported rates of undercontrolled behavior (such as aggression or disobedience) between American and Thai children on the Child Behavior Checklist (CBCL, Achenbach & Edelbrock, 1983). Normative use of behavior ratings across cultures may be misleading. Achenbach et al. (1990) compared CBCL scores for Puerto Rican and mainland American children and found significant mean differences on CBCL scores between groups. This was true for parent report, teacher

report, and self- report. The use of mainland American norms would have resulted in a 45.3% prevalence rate for psychopathology among Puerto Rican children (Bird et al., 1988).

Cross-cultural differences in professionals' assessment of behavior using behavior ratings also have been reported. Mann et al. (1992; Mueller et al., 1995) asked mental health professionals and teachers to rate videotaped vignettes of students' behavior using a scale of items derived from DSM-III-R criteria for ADHD, oppositional defiant disorder, and conduct disorder the Conners Abbreviated Teacher Rating Scale (ATRS, Conners, 1973) (one item was added by the research team). Mann et al. (1992) found that ratings from Chinese and Indonesian professionals were significantly higher than those of American and Japanese professionals. SonungaBarke, Minocha, Taylor, and Sandberg (1995) assessed the extent to which teachers' ratings of behavior corresponded to actometer readings and behavioral observations in two experiments. 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. Even a laboratory measure, such as the Gordon Diagnostic System (GDS; Gordon, 1982), can exhibit seriously different impact when used across cultural groups. Bauermeister et al. (1990) reported that using American norms with the GDS could seriously overidentify Puerto Rican children.

Only one of these studies (Sonuga-Barke et al., 1993) used a design in which raters from one culture rated children from a different cultural group. This allowed for an analysis of whether ratings scales (or raters) perform differently across different cultural groups. Relatively few studies of this nature are available. Langsdorf, Anderson, Waechter, Madrigal, and Juarez (1979) used the ATRS and found significant differences in the rates which different ethnic groups would be identified as ADHD. African American students were proportionally overidentified and Mexican American students were proportionally underidentified. Waechter, Anderson, Juarez, Langsdorf, and Madrigal (1979) found that African American students received higher ATRS scores than Caucasian or Mexican American students. Overrepresentation of African American children in school-identified samples and significant differences in mean rating scale scores also have been reported in other studies which utilized different identification procedures (e.g., Costello & Janiszewski, 1990; Jarvinen & Sprague, 1996; Lambert, Sandoval, & Sassone, 1978).

Although the results of these studies admit the possibility of cross-cultural differences, it is impossible to determine if differences are due to the use of the scale with a culturally different population or to a real difference in the base rate of ADHD-like behaviors across groups. Only one study, Jarvinen and Sprague (1996) has addressed this issue. They used the ADD-H Comprehensive Teacher's Rating Scale (ACTeRS; Ullmann, Sleator, & Sprague, 1991) to assess whether items functioned differentially across Caucasian, Mexican American, and African American subjects. The results suggested that, although some items were biased, there was no systematic pattern of item bias that would inflate the scores of minority students. However, this scale does not reflect current ADHD diagnostic criteria (APA, 1994).

For an assessment instrument to be equivalent across different cultural groups, it must demonstrate conceptual and normative equivalence (Marsella & Kameoka, 1989). Conceptual equivalence pertains to similarities in the conceptual meaning of the constructs used in assessment (Marsella & Kameoka, 1989). Normative equivalence implies that normative

standards developed for one culture are appropriate for another (Marsella & Kameoka, 1989). Bracken and Barona (1991) suggested procedures for the cross-cultural validation process that include (a) comparison of descriptive statistics, (b) analysis of internal consistency (Cronbach's alpha), (c) analysis of item intercorrelations, and (d) evidence of developmental age progression. Other techniques can also provide information on cross-cultural equivalence. The use of confirmatory factor analysis can provide an empirical assessment of the extent to which factor structures are congruent across different groups. Structural equation modeling (SEM) can be used to assess whether factor structures differ across groups and can provide simultaneous tests of the extent to which relations between factors, item loadings, and item error variances differ across groups. Techniques such as multidimensional scaling (MDS) and item bias analysis can determine whether differences in scale equivalence are present.

The purpose of this study is to explore the cross- cultural equivalence of the ADHD Rating Scale-TV School Version (DuPaul et al., 1997), across Caucasian and African American male, public school students. We first report descriptive statistics for each group. Next, we present the results of SEM analysis that compares the equivalence of the two-factor DSM-TV (APA, 1994) model of ADHD across groups. Finally, we report the results of a MDS analysis that assesses the extent to which item dissimilarities differ across groups.

METHOD

Participants

Students for this study were selected from a pool of 4,009 drawn from the normative sample of the ADHD-TV Rating Scale-School Version (DuPaul et al., 1997). Only male students for whom complete data were available were selected. The final sample consisted of 381 African American (AA) and 1,359 Caucasian American (CA) public school students ages 5 to 18. Although data were collected for both males and females, we limited the analysis to males because males typically receive higher scores on behavior rating scales than females (Barkley, 1990; Du- Paul et al., 1997), and thus there may be differences in a combined group which could affect analysis, and because ADHD is more frequently diagnosed in males than in females. Data on females will be reported in a separate study. Because the total behavior rating scores decrease as age levels increase, we tested for the possibility of an age by ethnicity dependency. The results $x^2(13) = 20.57$, $p = .08$, suggest that the sample is proportional across age and ethnicity and thus is appropriate for analysis.

Participating teachers were largely female (82%). Teachers were primarily Caucasian (93.4%). Other ethnic groups included African Americans (5.4%), Hispanics (0.7%), Native American (0.1%), Asian American (0.1%), and Other (0.4%). To check for teacher-related differences that could confound interpretation, we tested whether teachers' years of experience differed and whether there were differences in the number of CA and AA ratings by special education versus general education teachers. Mean years of experience (and standard deviations) for teachers rating CA and AA students were 14.87 (9.16) and 13.64 (9.20), respectively. There was a significant difference in teaching experience, $t(1767) = 2.36$, $p = 0.019$; however, the large sample makes it likely that even trivial differences will be significant. The effect size was small (0.13), suggesting this was not a meaningful factor. The majority of teachers (91.1%) were general educators; the remainder (8.9%) were special educators. No dependency was found between ethnicity and special education/general education status ($x^2(1) = 0.82$, $p = .36$).

Instrumentation

Teachers completed the ADHD Rating Scale-TV (School Version) (DuPaul et al., 1997; see Appendix), which consists of 18 items directly adapted from the ADHD symptom list as specified in the DSM-TV. Teachers were asked to rate the behavior of two randomly selected students from their class roster (e.g., third male and fifth female on the class roster). Ratings were completed between October and May in the 1994-1995 or 1995-1996 school years. Estimated return rates ranged from 50-95% *(M* = 85%) across school districts. lb minimize possible bias due to response set, Inattention (TA) symptoms were designated as odd-numbered items while Hyperactivity-Impulsivity (HT) symptoms were designated as even-numbered items. Teachers were instructed to select the single response for each item that best described the frequency of the specific behavior displayed by the target child over the past 6 months (or since the beginning of the school year). The frequency of each item or symptom was delineated on a 4-point Likert scale ranging from 0 *(never or rarely)* to 3 *(very often),* with higher scores indicative of greater ADHD-related behavior.

Analysis

Descriptive Data

Descriptive data analysis followed the guidelines suggested by Bracken and Barona (1991). We computed descriptive statistics for each group including frequency distributions of scores and mean scores by age level for both HT and IA factors. We also assessed the distributions of individual items across groups and computed reliability coefficients (Cronbach's alpha) for HT and TA factors as well as total score.

Structural Equation Modeling Analysis

We used structural equation modeling (SEM) (LTSREL 8) to compare the factor structure, factor correlations, item loadings, and item uniquenesses across groups following procedures suggested by Joreskog and Sorbom (1993) and Benson (1987). We tested the two-factor model of ADHD based on DSM-TV (APA, 1994) definition which conceptualizes ADHD as having HI and IA factors. This model was based on factor analysis conducted on the ADHD-TV rating Scale-School Version during the norming process and was shown to adequately model observed data (DuPaul et al., 1997). To test whether factor structure was invariant across groups we used a four-step procedure. First we compared the CA and AA groups with all parameters (i.e., item factor loadings, factor correlations, and item uniquenesses) constrained to be equal to those of the CA subgroup. This step provides a baseline estimate of model fit. Second, we computed separate estimates of item factor loadings across subgroups. Third, we computed separate estimates of factor loadings and factor correlations across subgroups. Finally, we computed separate estimates of factor loadings, factor correlations, and item uniqueness across subgroups. (An item's uniqueness is composed of two components, random measurement error and unique item variance or variance not shared with other scale items.) Thus at each step an additional parameter was freed. The results from Steps 2-4 were then compared to the results of the previous step to assess whether model fit improved by freeing a constraint (i.e., allowing separate estimates of parameters for each group).

To assess whether freeing a restraint improved the overall fit of the model, we used the chi-square difference test suggested by Bentler and Bonnett (1980):

$$\chi^2_{(\text{difference})} = \chi^2_{(\text{model 1})} - \chi^2_{(\text{model 2})} \text{ and } df = df_{(\text{model 1})} - df_{(\text{model 2})}$$

where chi-square for Model 1 represents the observed chi-square value for the more restrictive model, and the chi-square for Model 2 represents the observed chi-square for a model with one restriction freed (with corresponding degrees of freedom). All parameter estimates were performed using covariance matrices and generalized least squares (GLS) estimation. GLS estimation was used throughout, because preliminary analysis showed that item distributions were skewed and highly kurtotic in many instances; GLS estimation was indicated as it has been shown to perform well under conditions of extreme skewness and kurtosis (Chou & Bentler, 1995).

Multidimensional Scaling Analysis

MDS analysis is a useful tool for discovering and representing the underlying structures (termed dimensions) of psychological constructs (MacCallum, 1988). Raw data are converted into a matrix representing the degree of similarity (or dissimilarity) between all pairs of items. The process is analogous to a mileage chart where cities are listed in a row and column on the margins of the chart. Tndividual cells in the chart correspond distances between pairs of cities. This allows data to be represented as points within a spatial representation that correspond to items within a set (Young & Hamer, 1987). This spatial representation in turn allows comparisons of whether dimensions and item similarities vary across groups, and thus allows an assessment of whether the actual constructs measured differ across groups.

We used MDS (ALSCAL) to create a dissimilarity matrix of distances between items for the CA and AA groups, using Euclidean distances and a nonmetric model (because data were ordinal), to assess the pattern of differences among items across the CA and AA groups. Each group was scaled separately. We hypothesized that if the same construct was being assessed for both groups, only dimensions representing ADHD would be present. Any additional stable dimensions would indicate the possible presence of rater or method effects. lb determine the number of dimensions, we followed procedures suggested by Davison (1983). We first computed separate three-dimensional solutions for the CA and AA groups. We then plotted stress values and variance accounted for by dimensions (in a procedure analogous to a scree plot in factor analysis) and analyzed derived stimulus configuration plots to determine the most appropriate number of dimensions (Davison, 1983). Finally, we randomly split each group into two subgroups and performed an orthonormal rotation to maximum convergence on the derived stimulus configuration. We then computed coefficients of congruence for each dimension between subgroups and the entire CA and AA groups. This allowed for an assessment of the stability of possible dimensions which is necessary to assess whether a derived dimension is a chance occurrence (and actually represents "noise") or is a true dimension.

RESULTS

Descriptive Data

The mean scores for the AA group were significantly higher than the CA group for both factors. Because both groups exhibited significant heterogeneity of variance ($F = 25.65$, $p < .001$ for HT; $F = 8.67$, $p < .001$ for TA), we used unequal variances for *t* tests. For the HI factor, means (and standard deviations) for the AA and CA groups were 9.64 (8.38) and 6.35 (7.24), t(555.63) = -7.0%. For the TA factor, means (and standard deviations) for the AA and CA groups were

11.84 (8.31) and 8.56 (7.70), t(589.71) = -6.98, $p < .001$. Effect sizes for HT and TA factor differences (using pooled variance) were .43 and .41, respectively. Both groups showed evidence of developmental age progressions as indicated by the decline in scores as age level increased.

The score distributions differed markedly across the two groups. For the CA group, the distribution showed a pronounced positive skewness for both HI and IA factors. However, for the AA group, the distribution of scores tended toward the platykurtic (or flat) for both the HT and TA factors. As a result, for the AA group, proportionally more scores fell at the high end. Thus, the differences between the proportion of scores at the low and high end of the scales was much more pronounced for the CA sample than the AA. The difference in distributions resulted in different cut points if the scale were used for screening/diagnostic purposes. For the HT factor, the 90th, 95th, and 98th percentile scores for the CA group corresponded to scores of 18, 22, and 25. Tf these cut points were used with the AA group they would identify 19.7, 11.2, and 5.5%, respectively, of the AA sample. For the TA factor, the 90th, 95th, and 98th percentile scores for the CA group corresponded to scores of 20, 23, and 25. Tf these cut points were used with the AA group they would identify 20.1, 11.3, and 5.7%, respectively, of the AA sample. Differences in distributions also were evident at the item level as evidenced by differences in skewness and kurtosis across groups. Reliability was high for both groups. For both the AA and the CA group Cronbach's alpha for the HT, TA, and total scale were .95, .95, and .96, respectively.

SEM Analysis

Two separate exploratory factor analyses were conducted for the AA and CA groups, using principal axis extraction and oblique rotation. We used principal axis extraction and oblique rotation, as opposed to the more commonly used principal components extraction and varimax rotation (Reid, 1995), because the use of principal components extraction with varimax rotation can sometimes result in creation of potentially extraneous minor factors or splitting larger factors into a number of smaller factors (Gorsuch, 1983), and because previous research has shown that HT and TA factors are highly correlated (DuPaul et al., 1997). The results of the exploratory analyses indicated that a two-factor solution consistent with DSM-IV two-factor conceptualization fit observed data. Both groups exhibited identical structures, with odd-numbered items constituting an TA factor and even-numbered items constituting an HT factor. We then assessed the model fit across the AA and CA groups. Table T shows the results of the SEM analysis. In each test, the invariant model, where factor loadings, factor correlations, and item uniquenesses are constrained to be equal to those observed in the CA subgroup, served as a baseline estimate. In subsequent tests, each of these restrictions were relaxed in turn. The model fit improved significantly only when item uniquenesses were allowed to vary across groups as evidenced by the increase in goodness of fit index (GFT) and a significant decrease in the value of chi- square. This suggests that, although the factor structures and correlations are the same, allowing the values of individual item uniquenesses to differ across groups significantly improved model fit. However, the improvement was slight. Overall model fit across groups, assessed by examining fit at step four, where all parameters are freed, is equivocal. The value of the root mean square error of approximation (RMSEA) and the GFT suggest there is not a good fit. The RMSEA is greater than the .08 value which suggests a reasonable error of approximation (Joreskog & Sorbom, 1993), and the GFT is less than the .90 level suggested as indicative of good model fit (Joreskog & Sorbom, 1993). In contrast, both the standardized root mean residual (SRMR) and the parsimony normed fit index (PNFT) are at or near the level suggesting at least a

moderate fit. Thus, there appears to be at least some degree of difference in model fit across groups. Table TT shows the values of the item factor loadings, uniquenesses, and R^2 (the squared, multiple correlation of each item with the remaining items) for each group. As would be expected, the

Table I. Test of Model Fit Across Caucasian and African American Samples[a]

Model	df	χ^2	$df_{(df)}$	$\chi^2_{(dif)}$	GFI	RMSEA	SRMR	PNFI
Invariant	305	3088	—	—	.70	.101	0.11	0.99
+Loadings	287	3079	18	9	.70	.104	0.14	0.93
+Factor correlations	286	3079	1	0	.70	.104	0.14	0.93
+Uniqueness	268	2764	18	315[b]	.75	.103	0.13	0.87

[a]GFI = goodness of fit index, RMSEA = root mean square error of approximation, SRMR = standardized root mean residual, PNFI = parsimony normed fit index.
[b]$p < .01$.

Table II. Item Factor Loadings, Uniquenesses, and R^2

	Caucasian			African-American		
Item no.	Item loading	Uniqueness	R^2	Item loading	Uniqueness	R^2
1	.81	.18	.79	.80	.15	.81
2	.84	.16	.81	.83	.11	.86
3	.87	.10	.89	.86	.10	.88
4	.87	.15	.83	.87	.10	.88
5	.82	.25	.73	.77	.22	.73
6	.84	.14	.83	.87	.03	.96
7	.83	.12	.85	.78	.09	.88
8	.86	.19	.79	.82	.20	.77
9	.87	.11	.87	.87	.12	.86
10	.86	.13	.84	.87	.07	.92
11	.85	.13	.85	.80	.12	.84
12	.83	.15	.82	.83	.15	.82
13	.85	.13	.85	.85	.09	.89
14	.80	.12	.84	.81	.11	.86
15	.93	.09	.91	.94	.09	.91
16	.88	.06	.92	.87	.07	.92
17	.88	.10	.88	.86	.07	.91
18	.89	.07	.92	.88	.06	.93

values of the factor loadings are similar. In contrast, there are differences in item uniquenesses across groups. The differences tend toward higher values for the CA groups. Analysis of modification indices indicated that the greatest differences across groups, in terms of effects on model fit, were for items 2, 4, 6, 7, 10, and 13. For these items, uniquenesses for the AA group were lower than for the CA group and explained variance was higher. This suggests that these items had less unique variance for the AA group, and thus were more strongly related to other scale items for the AA group as opposed to the CA group.

MDS Analysis

Initial analysis of stress values (using Kruskal's stress formula 1) found stress levels for the CA group of .1721, .1057, and .0692, for the one to three dimensional solutions, respectively. For the AA group, stress levels for the one-, two-, and three-dimensional solutions were, .1793, .1269,

and .0799. Stress values for one-dimensional solutions were both above .15 suggesting that a one-dimensional solution did not adequately represent the data. Stress levels for the two-dimensional solution for both groups approximated the .10 level which is acceptable although not optimal. The three-dimensional solution did not re duce stress below the desired level of .05. Coefficients of congruence for the AA subgroups were .989, .905, and .760 for dimensions one to three, respectively. Coefficients of congruence for the CA subgroups were .991, .921, and .824 for dimensions one to three, respectively. The results of this process suggested that Dimensions 1 and 2 are quite stable, while Dimension 3 is much less stable. Therefore, a two-dimensional solution most parsimoniously represented the data and allowed for interpretable results. Coefficients of congruence for the two-dimensional solution across the CA and AA groups were .986 and .388. This suggests that only Dimension 1 is congruent across the two groups.

Figures 1 and 2 show the rotated stimulus coordinates for Dimension 1 and 2 for each group. In each figure, the central axis represents the 18 items of the scale. Dimension 1 clearly separates all odd- (Lk factor) and even-(HT factor) numbered items, and thus represents an ADHD dimension. Distances between items for the two groups are minimal; no items display marked separation. This indicates that, for this dimension, the groups are quite similar in terms of the pattern of dissimilarities between items. For Dimension 2, there are obvious disparities between the groups on many items suggesting a different pattern of dissimilarities for this dimension; however, there is no obvious pattern of differences.

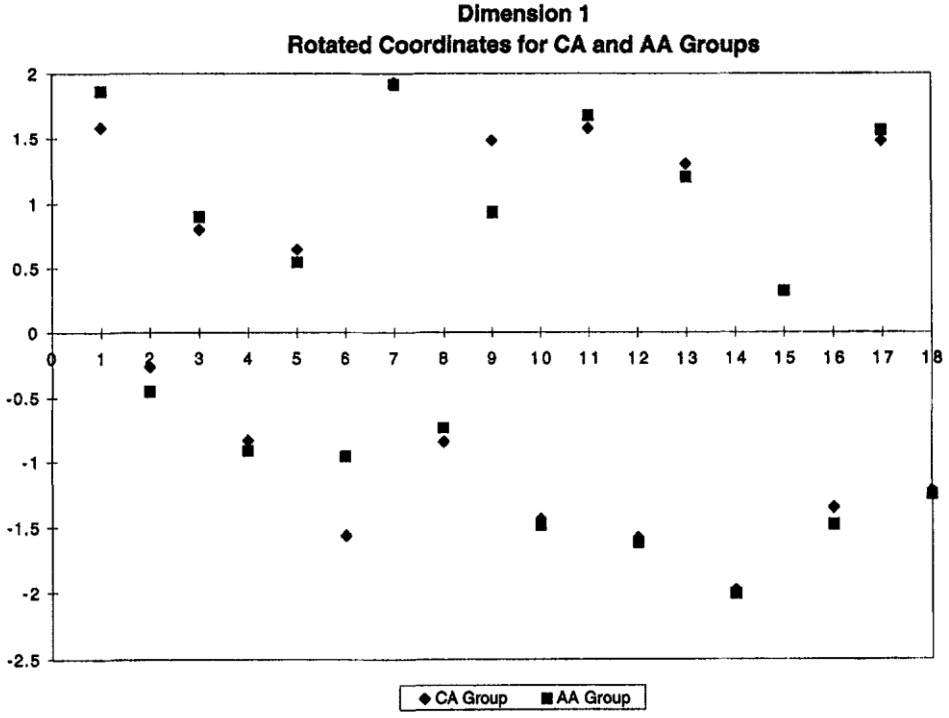

Fig. 1. Rotated stimulus coordinates for Dimension 1.

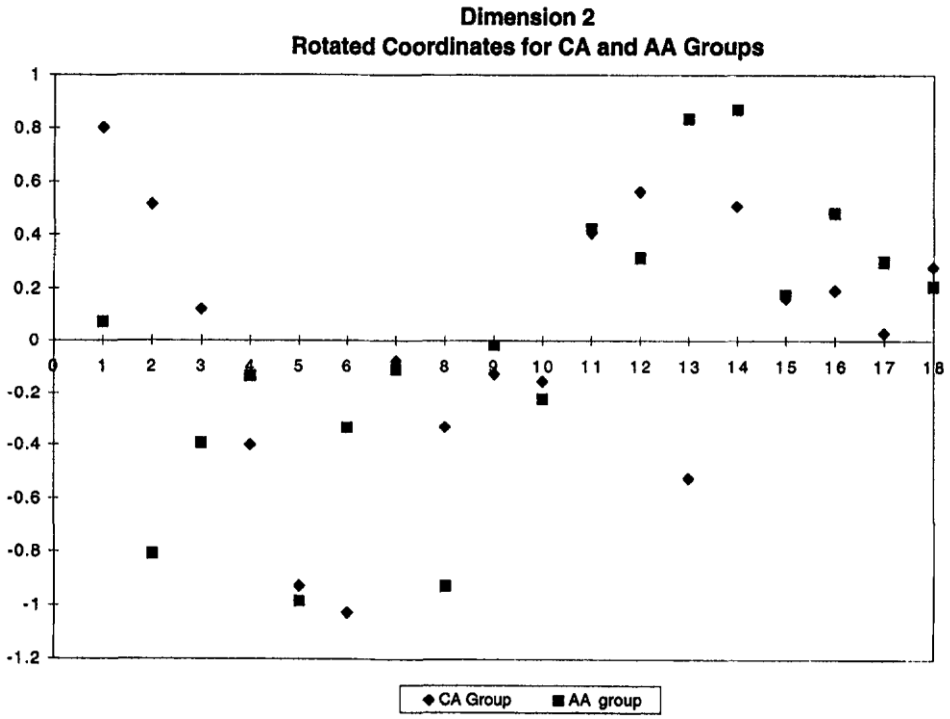

Fig. 2. Rotated stimulus coordinates for Dimension 2.

The greatest differences are for items 1, 2, 6, 8, and 13.

DISCUSSTON

The results of descriptive, SEM, and MDS analyses are consistent and suggestive of differences across the AA and CA groups. The analysis of descriptive data are consistent with previous research (e.g., Jarvinen & Sprague, 1995; Langsdorf et al., 1979; Waechter, et al., 1979) and suggest that there are consistent mean differences across groups. Moreover, there are significant differences in group variances and distinctly different distributions across groups for both the HT and IA factors and the individual items. As a result, if CA norms were used for AA students, approximately twice the number of AA students would screen positive on the HT and/or TA factors. Although these differences in descriptive statistics suggest a lack of cross-cultural equivalence, they are not sufficient. The differences could be due to real differences in actual behavior, instrument bias, or a combination of the two.

Results of SEM analysis suggest that the scale measures the ADHD construct somewhat differently across groups. Although the number of factors, factor correlations, and item loadings were equivalent across groups, there was a disparity due to the difference in item interrelations across groups as evidenced by the fact that different item uniquenesses were appropriate for the AA group. The results of the SEM analysis suggest that although the same underlying structure (i.e., HT and TA factors) is appropriate for both groups, the constructs themselves, although similar, are not identical. Thus the scale may not satisfy the conceptual equivalence requirement across groups. This suggests that at least some of the observed group differences are due to variations in the performance of the scale across groups as opposed to differences in actual behavior

exhibited. However, although there appear to be differences, the magnitude of the differences does not seem to be great for any single item. Rather, the difference appears to be the cumulative result of a number of small differences across groups. However, we again stress that although the differences were statistically significant they were relatively small.

Finally, the MDS analysis suggests that whereas the groups are similar in terms of the observed pattern of item dissimilarities for Dimension 1 (which we interpret to be an ADHD dimension), there are pronounced differences for some items in Dimension 2. This suggests that differences across groups are not primarily due to differences in the ADHD construct but rather to differences in Dimension 2. This could represent the possibility of a rater dimension which would suggest differences in teachers' perceptions of behavior between the CA and AA groups. This in turn suggests the possibility that some component of the scores of the AA and CA groups differed due to different interrelations among items. There is some congruence between the results of the SEM and MDS analyses. Both analyses suggest that the ADHD construct is similar across groups. Three of the items (2, 6, and 13) which exhibited differences across groups in the MDS analyses on Dimension 2 also exhibited differences in item uniquenesses; Item 6 also exhibited the greatest disparity on Dimension 1. This provides additional support for the notion that there is a difference in interrelations among items across groups.

The differences in item interrelations could be due to a halo effect. The lower unique variances observed in the AA group suggest that most items in the scale tend to covary consistently. Thus, if students in the AA group received high ratings on the HT subscale they would also tend to receive high ratings on the TA subscale. The presence of halo effects has been noted previously. Abikoff, Courtney, Pelham, and Koplewicz (1993) have reported that when teachers rate students with oppositional behaviors, halo effects, which serve to inflate ratings of ADHD-like behavior, are likely. Thus, for example if teachers tended to perceive African American students as oppositional, a halo effect would be likely.

The results of this study are consistent with those of Sonunga-Barke et al. (1993) as they suggest that factors other than student behavior may affect behavior ratings. However, it differs from the results reported by Jarvinen and Sprague (1995) in that it suggests that there may be a component of item scores which biases results of the AA group. There are a number of possible reasons for the disparity. First, the scales themselves are quite different. The ACTeRS contains fewer items in the HI and IA factors than the ADHD-TV RS (11 vs. 18). Ttem wording also differ between scales. Some ACTeRS items are quite global in nature (e.g., Works well independently) and would subsume a number of ADHD- TV RS items; other ACTORS items split items which are combined in the ADHD-TV RS. For example Item 2 on the ADHD-TV RS, "Fidgets with hands or feet or squirms in seat" are reflected in two separate ACTeRs items, Number 9 (Fidgety-hands always busy) and Number 11 (Restless-squirms in seat). Scoring also differs. Thus, the scales are not directly comparable. Second, there are distinct differences between the analyses. Jarvinen and Sprague's procedure assessed conditional probabilities across groups for each item after matching subjects on a criterion variable—ACTeRS total scale score. The use of the total ACTeRS scale score to equate individuals may also affect results, as the ACTeRS total score in-cludes two factors, Social Skill and Oppositional Behavior, which are not included in the DSM-TV diagnostic criteria. Additionally, the presence of halo effects due to oppositional behaviors

(Abikoff et al., 1993) may also have affected the process of equating individuals. In contrast, our analysis focused on the underlying structures which compose the scale.

We should caution that because we were not able to include socioeconomic status (SES), we cannot ascribe results entirely to cultural differences. When we analyzed ADHD-IV School Version scores and SES (for a subgroup of schools for which SES data were available). we found significant correlations. Thus, differences across groups may be due to differences in SES, cultural differences, or a combination of both. However, if SES differences are closely associated with ethnicity, this may be a distinction without a difference. We should also caution that, although there appear to be differences in the performance of the scale across groups we cannot assess the extent to which these differences might affect ratings. Taken as a whole, the results suggest that norms for the CA group may not be appropriate for the AA group. We further caution that, if a bias exists, even when AA students are proportionally represented in norm groups they are still more likely to screen positive for ADHD, due to the fact that their addition will have a relatively small effect on the entire distribution. For example, if the groups were combined, using the 90th, 95th, and 97th percentiles would result in 18.2, 9.4, and 5.5% of the AA group screening positive for the HI factor, and 17.7, 7.7, and 5.7% of the AA group screening positive for the TA factor.

In sum, the results of this study suggest the possibility that factors other than behavior may affect the results of behavior rating scales for AA students. The results of the study have two significant implications for assessment. First, because the scale reflects DSM-TV diagnostic criteria for ADHD, it suggests that there is the possibility that student ethnicity may affect the likelihood of a rater endorsing the presence of ADHD symptoms. One possible procedure to address this issue might be to compare scores to both a standardized norm group (i.e., a norm group composed of proportionally represented ethnic groups) and to a peer norm group, representing only that child's ethnic group. This would allow both an assessment of symptom severity in terms of the overall population, and an assessment in terms of a peer group. Second, the results suggest that there may be the possibility of halo effects for AA students. Thus, clinicians might expect to see a disproportionate number of AA children who would be diagnosed as ADHD combined type. This suggests that, for AA children, clinicians should rely more heavily on behavioral observations as opposed to behavior ratings.

This "halo" hypothesis could be tested by gathering teacher ratings on ADHD and oppositional behaviors and comparing them to observational data to determine if there is a negative halo effect for AA children. It would be also be of interest to assess the extent to which rater ethnicity (i.e., whether the teacher was AA or CA) affected ratings. Because of the small number of AA teachers in this study, we were not able to make any comparisons of this nature. Additionally, future researchers might address whether differences found in this study would occur across AA and CA parent groups.

In conclusion, we note that the results of this and previous studies argue for caution in the use and interpretation of rating scales with culturally different students. However, we caution that these results should be viewed as preliminary until they can be replicated across several more samples. Moreover, we caution that the methods we used in this study do not allow us to assess whether differences noted in the SEM and MDS analysis have a meaningful effect on diagnoses.

Additional experimental studies incorporating observational data and both AA and CA raters are needed to determine the practical impact of cultural differences on raters and rating scales scores.

ACKNOWLEDGMENT

Special thanks to Deborah Bandalos, Cal Gar- bin, and Melody Hertzog, University of Nebraska-Lincoln, for providing their assistance and insights during the preparation of this manuscript.

APPENDIX
ADHD RATING SCALE-IV-SCHOOL VERSION

Child's Age_____ Grade_____ Ethnic background_____

From your class roster select the second male child. Circle the one number that best describes this student's school behavior over the past 6 months (or since the beginning of the school year). Please complete all items.

	never or rarely	sometimes	often	very often
1. Fails to give close attention to details or makes careless mistakes in schoolwork.	0	1	2	3
2. Fidgets with hands or feet or squirms in seat.	0	1	2	3
3. Has difficulty sustaining attention in tasks or play activities.	0	1	2	3
4. Leaves seat in classroom or in other situations in which remaining seated is expected.	0	1	2	3
5. Does not seem to listen when spoken to directly.	0	1	2	3
6. Runs about or climbs excessively in situations in which it is inappropriate.	0	1	2	3
7. Does not follow through on instructions and fails to finish work.	0	1	2	3
8. Has difficulty playing or engaging in leisure activities quietly.	0	1	2	3
9. Has difficulty organizing tasks and activities.	0	1	2	3
10. Is "on the go" or acts as if "driven by a motor."	0	1	2	3
11. Avoids tasks (e.g., schoolwork, homework) that require sustained mental/effort.	0	1	2	3
12. Talks excessively.	0	1	2	3
13. Loses things necessary for tasks or activities.	0	1	2	3
14. Blurts out answers before questions have been completed.	0	1	2	3
15. Is easily distracted.	0	1	2	3
16. Has difficulty awaiting turn.	0	1	2	3
17. Is forgotful in daily activities.	0	1	2	3
18. Interrupts or intrudes on others.	0	1	2	3

REFERENCES

Abikoff, H., Courtney, M., Pelham, W., & Koplewicz, H. (1993). Teachers' ratings of disruptive behaviors: The influence of halo effects. *Journal of Abnormal Child Psychology, 21,* 519533.

Achenbach, T. M., & Edelbrock, C. (1983). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile.* Burlington: University of Vermont, Department of Psychiatry.

Achenbach, T. M., Bird, H. R., Canino, G., Phares, V., Gould, M. S., & Rubio-Stipec, M. (1990). Epidemiological comparisons of Puerto Rican and U.S. mainland children: Parent teacher and self-report. *Journal of the American Academy of Child and Adolescent Psychiatry, 29,* 84-93.

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 Education and American Life.* Washington, DC: Author.

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

Barkley, R. A. (1990). *Attention deficit hyperactivity disorder: A handbook for diagnosis and treatment.* New York: Guilford.
Bauermeister, J. J., Berrios, V., Jimenez, A. L., Acevedos, L., & Gordon, M. (1990). Some issues and instruments for the assessment of attention-deficit hyperactivity disorder in Puerto Rican children. *Journal of Clinical Child Psychology, 19,* 9-16.
Benson, J. (1987). Detecting item bias in affective scales. *Educational and Psychological Measurement, 47,* 55-67.
Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588.606.
Bird, H., Canino, G., Rubio-Stipec, M., Gould, M., Ribera, J., Sesman, M., Woodbury, M., Huertas-Goldman, S., Pagan, A., Sanchez-Lacay, A., & Moscoso, M. (1988). Estimates of the prevalence of childhood maladjustment in a community survey in Puerto Rico. *Archives of General Psychiatry, 45,* 11201126.
Bracken, B., & Barona, A. (1991). State of the art procedures for translating, validating, and using psychoeducational tests in cross-cultural assessment. *School Psychology International, 12,* 119-132.
Chinn, P. C., & Hughes, S. E. (1987). Representation of minority students in special education classes. *Remedial and Special Education,* 8(4), 41-46.
Chou, C., & Bentler, P. (1995). Estimates and tests in structural equation modeling. In R. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and practice* (pp. 37-55). Thousand Oaks, CA: Sage.
Conners, C. K. (1973). Rating scales for use in drug studies with children. *Pharmacology Bulletin* (special issue—Pharmacotherapy of Children) Vol. 24-29.
Costello, E. J., & Janiszewski, S. (1990). Who gets treated? Factors associated with referral in children with psychiatric disorders. *Acta Psychiatrica Scandinavica, 81,* 523-529.
Davison, M. (1983). *Multidimensional scaling.* New York: Wiley. Diana v. State Board of Education. Civ. No. C7037 RFP (N.D Cal. 1970, 1973).
Dunn, L. M. (1968). Special education for the mildly retarded: Is much of it justifiable? *Exceptional Children, 35,* 5-22. DuPaul, G. J., Power, T. J., Anastopoulos, A., Reid, R., McGoey,
E., & Ikeda, M. J. (in pres8). Teacher ratings of attentiondeficit/hyperactivity disorder symptoms: Factor structure, normative data, and psychometric properties. *Psychological Assessment.*
Epstein, M. A., Shaywitz, S. E., Shaywitz, B. A., & Woolston, J. (1991). The boundaries of attention deficit disorder. *Journal of Learning Disabilities, 24,* 78-86.
Gordon, M. (1982). *The Gordon diagnostic system.* DeWitt, NY: Gordon Systems.
Gorsuch, R. (1983). *Factor analysis.* Hillsdale, NJ: Erlbaum. Hunt, N., & Marshall, K. (1994). *Exceptional children and youth.* Boston: Houghton Mifflin.
Jarvinen, D. W., & Sprague, R. L. (1995). Using ACTeRS to screen minority children for ADHD: An examination of item bias. *Journal of Psychoeducational Assessment, 13, (Special Issue on ADHD),* 172-184.
Joreskog, K., & Sorbom, D. (1993). LISREL 8. Hillsdale, NJ: Erlbaum.
Lambert, N. M., Sandoval, J., & Sassone, D. (1978). Prevalence of hyperactivity in elementary school children as a function of social system definers. *American Journal of Orthopsychiatry, 48,* 446-463.

Langsdorf, R., Anderson, R. P., Waechter, D., Madrigal, J., & Juarez, L. (1979). Ethnicity, social class, and perception of hyperactivity. *Psychology in the Schools, 16,* 293-298.
Larry P. v Riles. C-71-2270-RFP (N. D. Cal. 1972) 495 F. Supp. 96 (N. D. Cal. 1979) Affr (9th Cir. 1984), 1983-84 EHLR DEC. 555:304.
Lerner, J., & Lerner, S. (1991). Attention deficit disorder: Issues and questions. *Focus on Exceptional Children,* 24(3), 1-17.
MacCallum, R. (1988). Multidimensional scaling. In J. R. Nesselrode & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 421-445). New York: Plenum Press.
Mann, E. M., Ikeda, Y., Mueller, C. W., Takahashi, A., Tao, K. T., Humris, E., Li, B. L., & Chin, D. (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 clinicians* (pp. 231-256). Washington DC: American Psychiatric Press.
Mercer, J. (1973). *Labeling the mentally retarded.* Berkley: University of California Press.
Mueller, C. W., Mann, E. M., Thanapum, S., Humris, E., Ikeda, Y., Takahashi, A., Tao, K. T., & Li, B. L. (1995). Teachers' ratings of disruptive behavior in five countries. *Journal of Clinical Child Psychology, 24,* 434-442.
O'Donnell, J. P., Stein, M. A., Machabanski, H., & Cress, J. N. (1982). Dimensions of behavior problems in Anglo-American and Mexican-American preschool children: A comparative study. *Journal of Consulting and Clinical Psychology, 50,* 643651.
Penning, N. (1990). Definitions of handicapping conditions expands . . .almost! *School Administrator,* 47, 31-32.
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 behavior rating scales. *School Psychology Review, 24,* 537-560.
Salvia, J., & Y8seldyke, J. (1988). *Assessment in special and remedial education* (4th ed.). Dallas, TX: Houghton Mifflin.
Sattler, J. (1988). *Assessment of Children* (3rd. ed.). San Diego, CA: Jerome Sattler.
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.
U.S. Congress. Committee on Education and Labor. (1990). *Hearings on the Reauthorization of the EHA Discretionary Programs.* Hearings, 101st Cong., 2d Sess. Feb. 21, 1990. Washington, DC: U.S. Government Printing Office.
Ullmann, R. K., Sleator, E. K., & Sprague, R. L. (1991). *ADD-H Comprehensive Teacher's Rating Scale-ACTeRS.* Champaign, IL: MetriTech.
Waechter, D., Anderson, R., Juarez, **L. J.,** Langsdorf, R., & Madrigal, J. F. (1979). Ethnic group, hyperkinesis, and modes of behavior. *Psychology in the Schools, 16,* 435-439.
Weisz, J. R., Suwanlert, S., Chaiyasit, W., Weiss, B., Achenbach, T. M., & Walter, B. (1987). Epidemiology of behavioral and emotional problems among Thai and American children: Parent reports for ages 6 to 11. *Journal of the American Academy of Child and Adolescent Psychiatry, 26,* 890-897.

Weisz, J. R., Suwanlert, S., Chaiyasit, W., Weiss, B., Achenbach, T. M., & Trevathan, D. (1989). Epidemiology of behavioral and emotional problems among Thai and American children: Teacher reports for ages 6 to 11. *Journal of Child Psychology and Psychiatry, 30,* 471-484.
Whalen, C. K. (1989). Attention deficit and hyperactivity disorders. In T. H. 011endick & M. Hersen (Eds.), *Handbook of child psychopathology* (2nd ed., pp. 131-169). New York: Plenum Press.
Young, F. W., & Hamer, R. M. (1987). *Multidimensional scaling: History, theory, and applications.* Hillsdale, NJ: Erlbaum.