Race and Ethnic Differences and Human Figure Drawings:
Clinical Utility of the DAP:SPED

Holly C. Matto
School of Social Welfare, University at Albany, State University of New York

Jack A. Naglieri
Center for Cognitive Development, George Mason University

This study examined race and ethnic differences on the Draw A Person: Screening Procedure for Emotional Disturbance (DAP:SPED; Naglieri, McNeish, & Bardos, 1991) for youths 6 though 17 years of age for 2 matched samples. Samples were drawn from the DAP:SPED nationally representative standardization sample and matched on gender, grade, and school classroom. No statistically significant differences were found for big figure, small figure, or shading item composites. A statistically significant difference was found between Black–White pairs on figure omissions but showed a small effect size (d value = .25). Further, no statistically significant differences were found between the DAP:SPED Total T scores for Black and White youth (M = 47.67, SD = 10.09; N =138) or Hispanic and White youth (M = 48.20, SD = 9.56; N = 59), showing very small effect sizes. In addition, equivalence testing showed similarities across race and ethnic pairs for all composites and DAP:SPED total score, lending preliminary support to the DAP:SPED’s clinical utility as a measure that yields similar scores across these groups.

Use of human-figure drawings remains a controversial topic and an example of the disparities between research and practice. Although human-figure drawing is among the most widely used psychological tests by clinicians (Camara, Nathan, & Puente, 2000), the validity of the technique has been questioned for some time (Roback, 1968; Swensen, 1957, 1968). Recently, researchers have again cautioned against the use of most projective techniques, including human-figure drawing tests. For example, Garb, Wood, Lilienfeld, and Nezworski (2002) have been most critical of the approach, stating that human-figure drawings significantly lack validity.

Naglieri, McNeish, and Bardos (1991) argued that human-figure drawing systems have low reliability and validity for several reasons. First, the scoring rules are often ambiguous, leading to low interrater as well as intrarater reliability. Second, the lack of an objective standardized scoring system as well as a nationally representative normative sample for the calculation of standard scores have resulted in poor psychometric qualities, which limit clinical interpretation of test results. Third, although interpretation of individual signs based on specific aspects of the human-figure drawing lack both validity and reliability, a composite or global system could be more effective. The approach that Naglieri et al. used in the development of their Draw A Person: Screening Procedure for Emotional Disturbance (DAP:SPED) test substantially improves on these earlier psychometric deficits and, as such, has led some researchers to conclude that it is the “most psychometrically advanced of the human-figure drawing instruments” (Matto, 2002, p. 221).

Researchers have found, for example, that the DAP:SPED has validity for identification of youths with emotional or behavioral disorders. In one study, the DAP:SPED was found to successfully differentiate conduct and oppositional defiant disorder and control group youths (Naglieri & Pfeiffer, 1992). Similarly, special-education students with identified emotional disturbance were differentiated from regular-education students (McNeish & Naglieri, 1993) using this scoring system.

In a recent study, Matto (2002) found that DAP:SPED scores were related to internalized behavioral disturbance in youths receiving clinical services through outpatient or residential treatment centers and offered predictive ability above and beyond a standard paper-and-pencil measure of internalizing disturbance. However, the DAP:SPED did not show significant predictive ability for externalizing disturbance above and beyond a standard paper-and-pencil measure of ex-
ternalizing disturbance in this study. Hit rates or other indicators of magnitude were not reported by these authors. In a dissertation study, Matavich (1998) examined the DAP:SPED scores of delinquent adolescents with emotional or behavioral disturbance who were placed in highly restricted educational or correctional settings (N = 154). Results showed the delinquent group mean score was significantly higher than a matched control group (N = 154) with an effect size d value of 1.6.

The accumulation of evidence from these validity studies on the DAP:SPED has led researchers critical of projective tests to conclude that “there is some evidence that the DAP:SPED (Naglieri et al, 1991) can be used to screen for mental disorders among children” (Garb et al., 2002, p. 459). However, the differences found in the existing literature regarding the nature of emotional and behavioral disturbance (internalizing or externalizing) that the DAP:SPED predicts indicates that more validation work is needed to refine assessment understanding along these behavioral dimensions. Additional study of this approach, therefore, is warranted.

Another important area of further study is examining the measure’s clinical utility across different race and ethnic groups. The potential differential utility of measurement instruments across groups has been of considerable interest to researchers who have developed tests for use with diverse groups and for clinicians who administer, interpret, and make decisions based, in part, on such test results. Studies that ask the question of whether race-based test differences exist are critical for establishing the clinical utility of an instrument and for expanding clinicians’ understanding of a test’s assessment range. This study did not directly test the different negative social consequences related to test application but examined the question of potential construct-irrelevant variance related to race or ethnicity.

Although race and ethnic differences are often studied for intelligence tests (e.g., Brody, 1992; Naglieri & Ronning, 2000), such comparisons are equally important for projective tests, particularly for tests that have the potential to influence the number of minority youths who are placed in special-education programs (Messick, 1995). For these reasons, examination of the differences between race and ethnic groups on a scoring system such as the DAP:SPED is important, particularly when minority overrepresentation in special-education programs is of concern (e.g., Oswald, Coutinho, Best, & Singh, 1999).

Messick (1995) argued that an instrument’s clinical utility must be questioned if the scores it yields contribute to problems such as overrepresentation of minorities in special-education classes for children with emotional or behavioral problems (Oswald et al., 1999; Reschly & Bersoff, 1999). Although some psycho-metric experts reject the use of mean score differences as evidence of test bias (see Reynolds & Kaiser, 1990), mean score differences among racial groups can be viewed as an indication of test bias if construct-irrelevant variance systematically caused the differences (Messick, 1995). Messick suggested that social value consequences related to construct-irrelevant variance (e.g., bias attributed to race that affects how scores are interpreted and acted on) is an important component of traditional validity testing. This study contributes to preliminary exploration of the DAP:SPED’s potential for score differences by racial and ethnic group membership.

Equivalence tests, along with traditional significance tests, can provide a more comprehensive understanding of equivalence across groups related to a meaningful clinical criterion, rather than as difference from zero. Current clinical research often includes effect sizes and equivalence testing, in addition to traditional significance testing, in presentation of cross-group results (Jaccard & Guilamo-Ramos, 2002; Pina, Silverman, Fuentes, Kurtines, & Weems, 2003; Rogers, Howard, & Vessey, 1993).

Naglieri et al. (1991) performed a limited examination of race and ethnic differences on the DAP:SPED, but their analytic methods did not examine composite scores across race and ethnic groups or include equivalence testing procedures. Specifically, prior work did not match race or ethnic pairs on important demographic variables or examine specific DAP:SPED item composites, such as shading of the figure, that might represent bias at the item level and contribute to overall differences. This study was conducted to address these limitations. The aims of this study were twofold.

The first study aim was to compare total DAP:SPED standard scores for Black and White samples, as well as Hispanic and White samples, matched on demographic variables. Matching procedures were used to select race pairs that were comparable on other variables, such as grade, gender, and school classroom, that could potentially influence DAP:SPED scores and be attributed to race differences. However, in creating matched race pairs we recognized that the procedure may have eliminated other “true” differences related to race not accounted for in the study.

The second study aim was to examine these race and ethnic pairs for differences on four item composites derived from the human-figure drawing test that could contribute to total score differences.

Method

The two samples included in this study consisted of Black–White (total N = 276; n = 138 for each race group in the pair) and Hispanic–White (total N = 118; n = 59 for each race group in the pair) children and ado-
lescents matched on gender, grade, and school classroom and drawn from a nationally representative standardization sample (N = 2,355) that mirrored the U.S. population. Data collection was conducted by a private company that followed appropriate procedures and federal regulations for use of human participants. This research was exempt from Institutional Review Board approval because it involved the use of educational tests in typical educational settings and the information was recorded in such a manner that student confidentiality was assured. All data collection procedures used were approved by appropriate members of the company.

All youths self-identified as White (European American), Hispanic (Spanish-speaking and from Hispanic culture), or Black (African American). All youths in the Hispanic group spoke Spanish and none of the youths in the White group spoke Spanish. The average age of the Hispanic–White paired sample was 11.80 years old (SD = 3.14, range = 6.03 to 17.36 years) and the average age of the Black–White paired sample was 11.30 years old (SD = 3.62, range = 5.05 to 17.89 years). There were 45.8% (n = 54) boys and 54.2% (n = 64) girls in the Hispanic–White matched samples and 47.1% (n = 130) boys and 52.9% (n = 146) girls in the Black–White matched samples. This study did not include additional standardized measures or administrative reports of youths’ emotional and behavioral problems that could have been used to assess potential differences between the race pairs on reported behavioral problems such as classroom disruption. However, race pairs were matched by classroom to create pairs that experienced similar classroom environments.

The DAP:SPED was developed using an actuarial approach to identification of individuals with emotional problems. The scoring system is an objective approach to determine the frequency with which unusual items that may be considered indicators of emotional problems occur in nonclinical versus clinical populations. The scoring system requires the production of three human-figure drawings (man, woman, self), which are scored using 55 carefully defined items. Examples of items include failed integration of body parts, presence of fists or talons, aggressive symbols, hands cut off, eyes omitted. The system is normed on 2,355 youths ages 6 to 17 years who are representative of the U.S. population on a number of important demographic variables. According to Naglieri et al. (1991), the sample represents the U.S. population in terms of age, gender, geographic region, race, ethnicity, and socioeconomic status. Rather than focus on individual interpretation of items, the DAP:SPED is a global system that yields a T score (M = 50, SD = 10), with separate T score norms for gender and age. The higher the score, the more likely the youth’s drawings are like those produced by youths who have demonstrated emotional or behavioral problems. Cronbach’s alpha internal reliability estimate of the DAP:SPED is as follows: ages 6 to 8 years, .76; ages 9 to 12 years, .77; and ages 13 to 17 years, .71.

The Matrix Analogies Test–Short Form (MAT–SF; Naglieri, 1985) was used as a nonverbal measure of general ability to control for possible differences in ability and was administered in English to all youth at the time of DAP:SPED standardization. The group-administered MAT–SF is composed of 34 progressive matrix type items standardized on a sample of 4,750 youths ages 5 through 17 years who represent the U.S. population on a number of key demographic variables. The MAT–SF has been found to possess good reliability (.83 median internal consistency, Cronbach’s alpha estimate). Validity of the test has been demonstrated by strong correlations with other nonverbal tests as well as achievement test scores (Naglieri, 1985). Additionally, the MAT–SF and its revision, the Naglieri Nonverbal Ability Test (Naglieri, 1997), has been shown to yield small mean score differences between White and minority groups (Naglieri, 1985; Naglieri & Ronning, 2000).

The samples’ MAT–SF (Naglieri, 1985) mean raw scores are as follows: Hispanic (M = 25.23, SD = 6.60) and White (M = 28.02, SD = 3.47) samples differed somewhat—t = 2.45, p < .05, d value was moderate (.53)—but the Black (M = 22.63, SD = 6.77) and White (M = 24.32, SD = 5.71) samples were similar—t = 1.82, p > .05, d value was small (.27).

Results

Separate analysis of variance procedures were conducted to examine Hispanic–White and Black–White race differences on DAP:SPED item composites and total T scores. To examine the question of whether minority youth might be disproportionately represented on item composites such as shading, item composites were analyzed along with the total T scores. Practitioners who use the DAP:SPED have consistently argued that this is a question that needs answered and that research has not yet examined it. For example, many racial and cultural groups may use figure shading to more accurately represent skin tone and may do so more frequently than nonminority youth. Given that the presence of shading is an indicator that increases a youth’s behavioral disturbance score in the total DAP:SPED calculations, minority youth may show systematically higher behavioral disturbance scores as a function of such test bias rather than reflecting true behavioral functioning.

Other human-figure drawing studies have found shading, omissions, and figure size to be significantly related to mental health problems (e.g., poor self-esteem, hostility; Burgess & Hartman, 1993; Burgess,
McCausland, & Wolbert, 1981; Hibbard & Hartman, 1990; Holmes & Wiederhold, 1982), although conventional practice is to make assessments from total human-figure drawing scores rather than at the item or composite level.

In addition, practicing clinicians using the DAP:SPED have anecdotally questioned the implicit assumption that these human-figure drawing constructs (shading, omissions, and figure size) are related to emotional and behavioral disturbance equally for all racial and ethnic groups. To date there has not been any DAP:SPED study to test these instrument-utility questions. Therefore, the four item composites included in this analyses were large figure (Items 1 and 3 summed); small figure (Items 2 and 4 summed); figure omissions, to include head omissions such as eyes or nose as well as body omissions such as feet or arms (Items 19 to 28, 45, 46 summed); and shading (Items 30 to 33 summed) summed across man, woman, and self drawings. These were included as the dependent variables of interest in the analyses.

No significant differences were found between the DAP:SPED Total T score means for the Hispanic (M = 49.02, SD = 10.34) and White (M = 48.20, SD = 9.56), F(1, 116) = .197, p = .658, or Black (M = 47.99, SD = 10.88) and White (M = 47.67, SD = 10.09), F(1, 274) = .067, p = .796, matched samples, with reported effect size d values for the two matched samples of .08 and .03, respectively. Similarly, no significant differences were found between Hispanic and White or Black and White matched samples on the basis of large figure, F(1, 116) = .738, p = .392, d value = .16; F(1, 274) = .000, p = 1.000, d value = .00, or small figure, F(1, 116) = .144, p = .705, d value = .07; F(1, 274) = .185, p = .667, d value = .05.

No significant differences were found between Hispanic and White or Black and White matched samples on the basis of shading item composites, F(1, 116) = .014, p = .904, d value = .02; F(1, 274) = 2.852, p = .092, d value = .21. Further, no statistically significant difference was found between Hispanic–White youth on figure omissions, F(1, 116) = .869, p = .353, d value = .17. However, a statistically significant difference was found on figure omissions for the Black–White sample (Items 24 to 28, 45, 46 summed across man, woman, and self drawings), with Black youth showing a higher mean score as compared to White youth, F(1, 274) = 4.175, p = .042, d value = .25. Additional analyses that partitioned the total figure omission variance at the item level showed the overall effect not to be statistically significant, F(1, 274) = 1.379, p = .223, suggesting that statistically significant single omission items cannot be identified in creating a meaningful difference between these race pairs. Along these lines, no statistically significant differences were found between either set of matched pairs on total DAP:SPED standard score, indicating that the significant difference in item composite found for figure omissions on the DAP:SPED does not seem to contribute to an overall total score differentiation between Black–White race pairs.

To further examine the utility of the DAP:SPED in assessment with minority youth, equivalence testing using the confidence interval approach (Rogers et al., 1993; Westlake, 1981) was applied (see Table 1). A 10% equivalence criterion was used for total DAP T scores for each set of race pairs. As Table 1 shows, the 90% confidence interval for both sets of race pairs is included within its respective equivalence interval, demonstrating clinical equivalence across Hispan-

<table>
<thead>
<tr>
<th>Table 1. Equivalence Testing Using Confidence Interval Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total DAP:SPED T Score</strong></td>
</tr>
<tr>
<td>Hispanic–White</td>
</tr>
<tr>
<td>Black–White</td>
</tr>
<tr>
<td><strong>Large Figure</strong></td>
</tr>
<tr>
<td>Hispanic–White</td>
</tr>
<tr>
<td>Black–White</td>
</tr>
<tr>
<td><strong>Small Figure</strong></td>
</tr>
<tr>
<td>Hispanic–White</td>
</tr>
<tr>
<td>Black–White</td>
</tr>
<tr>
<td><strong>Shading</strong></td>
</tr>
<tr>
<td>Hispanic–White</td>
</tr>
<tr>
<td>Black–White</td>
</tr>
<tr>
<td><strong>Omissions</strong></td>
</tr>
<tr>
<td>Hispanic–White</td>
</tr>
<tr>
<td>Black–White</td>
</tr>
</tbody>
</table>

Note: CI = confidence interval; LCL = Lower confidence limit; UCL = upper confidence limit; DAP:SPED = Draw A Person Screening Procedure for Emotional Disturbance.
ic–White and Black–White race pairs on total DAP T scores (Rogers et al., 1993).

An equivalence criterion of $\pm 0.468$ of one raw score point was used for both the large and small item composites, as this fraction represents the estimate for clinical significance as scaled to these two 6-item composites, using the clinical significance scoring guidelines provided in the DAP:SPED manual. Specifically, this fraction represents 7.8% of the 6-item total composite necessary to conclude the lowest threshold for inferring clinical significance and referring a youth for further evaluation. An equivalence criterion of $\pm 0.936$ of one raw score point was used for the 12-item shading composite and $\pm 7.02$ raw score points was used for the 36-item omissions composite, using this same clinical significance rationale. These calculations were used to construct the equivalency intervals, based on the White group means, and are shown in Table 1. As the results show, all 90% confidence intervals are contained within the equivalency intervals constructed based on clinical cut-off criteria, suggesting equivalence across race groups.

### Discussion

In summary, no statistically significant differences in the item composites tested (i.e., large figure items, small figure items, or shading) were found for the race and ethnic comparisons. Although an overall statistically significant difference was found for figure omissions between Black–White race pairs, subsequent analyses that partitioned the variance at the item level, showed no statistically significant difference between race pairs. Further, DAP:SPED total T scores did not differ significantly for the Black–White or Hispanic–White matched groups. Equivalence testing suggested both sets of youth pairs (Hispanic–White and Black–White) were equivalent across DAP:SPED item composites and DAP:SPED total score. These preliminary findings suggest that the race or ethnicity of the youth may have little influence on DAP:SPED T scores, showing lack of support for the presence of race or ethnicity-related construct-irrelevant variance. Certainly more studies will need to be conducted to build evidence of whether clinicians can, therefore, use this method with the expectation that it will not likely exacerbate problems of overrepresentation of minority youth in classes with those with emotional or behavioral disorders, as these preliminary findings would suggest.

This study contributes a new dimension to understanding the clinical assessment utility of the DAP:SPED, offering new evidence to a growing body of research that demonstrates practitioners can use this modern approach to projective testing in a way that is empirically supported. Other studies have examined the DAP:SPED’s domain-specific utility, suggesting that it is particularly successful in measuring internalizing rather than externalizing behavioral disturbance. This study examined a dimension of its culture-specific utility, offering preliminary evidence that the DAP:SPED may be appropriately applied across race and ethnic groups.

Further research should be conducted to examine larger, more diverse, and more culturally specific samples (e.g., different race and ethnic groups, bilingual children, cross-cultural populations with specific acculturation measures employed) and groups with different levels of socioeconomic status. An important limitation of this study is the absence of country of origin data for the youth. New studies should include this level of detail to examine these culturally specific research questions.

In addition, new studies that include standardized emotional or behavioral measures in the matching procedures across race pairs and that include cross-culturally validated ability measures need to be incorporated into future efforts to broaden understanding of the clinical utility of the DAP:SPED across race and ethnic groups. Finally, prospective studies that directly examine the predictive validity of the DAP:SPED should be the focus of new research agendas.

### References


Received February 26, 2004
Accepted January 26, 2005