

A Remedy for the Black-White Test-Score Disparity

Janet E. Helms
Boston College

Sackett, Schmitt, Ellingson, and Kabin (April 2001) analyzed the effectiveness of strategies for reducing the disparities in average scores on high-stakes tests of cognitive abilities (CATs) of (especially) African or Black and Latino and Latina Americans as compared with White Americans. They argued that decision makers in the domains of education, employment, and licensure and certification are becoming increasingly dependent on test scores as the primary criteria for making high-stakes decisions. Consequently, these two socio-racial groups, as well as Native and Asian Americans (with respect to tests of verbal skills), who are already underrepresented in many selective educational institutions and professions, may disappear from them entirely if the disparities in test scores cannot be eliminated or rendered meaningless for making high-stakes decisions involving them. Sackett et al.'s proposed solutions to the problem are to either "dumb down" (i.e., remove cognitive content of) the tests or alter the testing process so that it appears to be fair to Black and Hispanic test takers, even if, as the authors' analysis suggested, it is not.

In what seems to be an effort to prove that the test performance disparities between groups reflect actual irremediable cognitive deficiencies of the adversely affected test takers, Sackett et al. (2001) cited DeShon, Smith, Chan, and Schmitt's (1998) "unique study" (Sackett et al., 2001, p. 309) as disproving a "social relations and social context" (p. 309) argument, which they misattributed to me (Helms, 1992). I did not recommend that CAT items be modified to include social content. Most CATs already include such content. Instead, I discussed the absence of empirical evidence that CATs are culturally equivalent for African American test takers and proposed a strategy for quantifying the effects of racial and cultural variables on African, Latino and Latina, Asian, and Native American (ALANA) test takers' CAT scores. DeShon et al. allegedly collected the type of cultural data (e.g., racial identity attitudes) that could be used for trying the strategy but did not analyze it appropriately.

In Helms (1992), cultural equivalence was generally defined as the extent to which CATs measure the same intended constructs (e.g., knowledge, ability, skills) for (espe-

cially) ALANAs as compared with White test takers rather than some unmeasured constructs related to differential racial and/or cultural socialization. I discussed the potential effects on the construct validity of several types of equivalence. Also, I advised that one risks committing a cultural equivalence fallacy when one infers that high-stakes tests necessarily measure the same constructs (e.g., knowledge, ability) across groups without measuring and ruling out racial or cultural psychological variables as viable alternative explanations for the between-groups differences in CAT scores.

Of the various types of cultural equivalence fallacies that one might commit, (lack of) psychometric equivalence is most relevant to Sackett et al.'s (2001) analysis. *Psychometric equivalence* is defined as the extent to which a test measures the same construct equivalently across groups. Sackett et al.'s analysis is predicated on the premise that "well-developed tests in these domains [knowledge, skill, ability, and achievement] are valid for their intended purpose" (p. 302). However, valid tests are not necessarily culturally equivalent. That is, test scores may correlate in the same manner with criteria across groups if the test assesses some cultural or racial construct other than or in addition to the intended cognitive construct (Linn & Werts, 1971).

Sackett et al. (2001) marshaled an array of mostly secondary sources in support of their inference that high-stakes tests validly measure the cognitive abilities and skills of Black and Hispanic test takers. Yet virtually none of their cited sources investigated racial or cultural variables (other than demographic categories) as possible differential influences on the validity of ALANAs' test scores. Sackett et al.'s contention that CATs measure the same constructs for ALANAs as for Whites implied that the testing process is equivalent, but most of the validity research that they cited proceeds from the assumption or knowledge that the tests are not psychometrically equivalent (e.g., they have different between-groups means). This research investigated various strategies for gauging the level of adverse impact (e.g., cultural bias) on the ALANA groups of using tests to make decisions affecting them rather than the effects of culture on their test scores.

Helms (1992) hypothesized that cultural equivalence can be assessed by identifying and assessing the relations of emic racial or cultural psychological constructs to the ALANA group members' high-stakes test scores. *Emic* refers to constructs or processes specific to the racial or cultural socialization experiences of the lower scoring group. If test scores are correlated with emic variables, then the test is not culturally equiva-

lent and is probably not valid for its intended purpose with respect to that group—unless assessment of such cultural factors is the intended purpose of the test or test user.

Darlington (1971, p. 75), among others, offered a definition of *cultural fairness* that can be used to study cultural equivalence empirically. His definition of a culture-fair (i.e., equivalent) test requires that the scores on the cultural measure (C) and test (X) must be uncorrelated in a subset of people with the same criterion (Y) scores (i.e., $r_{XC.Y} = 0$). Generalizing from Darlington's definition, Y can be used to assess two different aspects of equivalence. If Y refers to the criterion (e.g., college grade point average; GPA) one intends to predict from test scores (e.g., on the Scholastic Assessment Test; SAT), then a significant correlation between test scores and the cultural variable (with the criterion controlled) would be evidence of cultural contamination in the test beyond what is relevant to predicting the criterion. This type of test-culture association would call into question Sackett et al.'s (2001) conclusions concerning comparable predictive validity of the tests. If Y is an alternative nontest measure (e.g., high school GPA) of the construct (e.g., cognitive ability) that the test is intended to assess, then a significant correlation between test scores and the cultural measure (when the nontest measure is controlled) would call into question Sackett et al.'s inferences with respect to construct validity (i.e., that the tests necessarily measure ALANAs' cognitive knowledge, ability, or skills).

In Table 1, construct and criterion equivalence analyses for the Black first-year students in Bradby, Helms, and Lyons's (2001)

sample, using the proposed definitions, are summarized. The emic cultural variable is Helms's (1990) Immersion–Emersion racial identity schema (e.g., idealization of one's Blackness and Black culture), high school GPA (HSGPA) operationally defines academic ability (i.e., the construct), and end-of-college or final undergraduate GPA (UGPA) is the criterion. The zero-order correlations (i.e., validity coefficients) are in the expected direction and are as high as or higher than those found in the validity literature that Sackett et al. (2001) cited (e.g., Wilson, 1981). Yet a lack of construct and cultural equivalence is suggested by the significant negative correlations between SAT-Verbal (SAT-V) and SAT-Mathematics (SAT-M) scores and culture when HSGPA or UGPA is controlled. That is, higher levels of Black idealization are associated with lower SAT scores, and, conversely, higher SAT scores are related to lower levels of Black idealization.

The regression equation for predicting SAT-V scores from HSGPA alone using the data in Table 1 is $SAT-V = 45.105HSGPA + 257.78$ (Equation 1). The multiple regression equation for predicting them from HSGPA and Immersion–Emersion racial identity schema (i.e., culture or C) is $SAT-V = -2.574C + 35.708HSGPA + 414.49$ (Equation 2). Using Equation 1, where culture is not measured, students with a 4.0 HSGPA are predicted to obtain SAT-V scores of 438. By Equation 2, the same students, with culture measured, are predicted to obtain SAT-V scores of 557, if they have zero levels of Black culture (as is typical of Whites). However, students with a perfect GPA and only an average amount of Black idealization (i.e., T score = 50) are predicted to obtain SAT-V

scores of 428, a slightly lower score than is predicted if culture is not measured at all (as is typically the case). The difference between the students' not having any culture (Equation 2) and the researchers' not measuring it (Equation 1) is 119 SAT-V points in favor of the cultureless Black students. The difference between not having any Black culture and having an average amount of it is 129 SAT-V points favoring the students without culture. Thus, culture potentially suppresses Black students' verbal test scores by more than one standard deviation, which exceeds the amount of disparity in ALANA–White test scores.

Space constraints preclude an extensive discussion of the potential suppressive effects of having culture on Black students' SAT scores, but generally both the SAT-V and the SAT-M test scores of brighter students (i.e., those having a HSGPA above the referent group average) can be expected to decrease by more than one standard deviation as levels of Black idealization increase. Thus, the remedy for improving Black students' test scores is obvious: Use educational interventions to deliberately destroy the positive racial identity development of bright Black students while they are in high school rather than leaving such destruction to chance. Alternatively, White students could be socialized to develop an idealized Black racial identity with the expectation that just being positively Black identified will cost them at least one standard deviation of (White) cultural advantage per subtest. If the relationships among variables discussed here generalize to other samples and emic racial or cultural psychological variables, then either of these proposed remedies, alone or combined with any of the various culturally sensitive remedies that Sackett et al. (2001) discarded, should eliminate the Black–White "valid" test-score disparity quite nicely.

REFERENCES

- Bradby, D., Helms, J. E., & Lyons, H. Z. (2001). *Are Black racial identity schemas better predictors of academic achievement than Scholastic Aptitude Test (SAT) scores?* Manuscript submitted for publication.
- Darlington, R. B. (1971). Another look at "cultural fairness." *Journal of Educational Measurement*, 8, 71–82.
- DeShon, R. P., Smith, M., Chan, D., & Schmitt, N. (1998). Can adverse impact on cognitive ability and personality tests be reduced by presenting problems in a social context? *Journal of Applied Psychology*, 83, 438–451.
- Helms, J. E. (1990). *Black and White racial identity: Theory, research, and practice*. Westport, CT: Greenwood Press.
- Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cogni-

Table 1

Correlations Between Scholastic Assessment Test (SAT) Scores and Culture With College or High School Grade Point Averages (GPAs) Controlled

Variable	Zero-order correlation			Culture correlation		M	SD
	UGPA	HSGPA	Culture	XC.Y(U)	XC.Y(H)		
SAT-V	.35	.24	-.33	-.23	-.25	390.84	88.33
N	107	104	119	106	114	119	
p	.0001	.014	.0001	.015	.006		
SAT-M	.22	.43	-.26	-.26	-.22	432.62	90.61
N	105	114	114	104	113	107	
p	.02	.0001	.005	.005	.016		
Culture	-.08	-.165					
N	107	107					
p	.42	.089	50.11				
M	2.56	2.95	10.33				
SD	.34	.47					

Note. Sample sizes vary because of missing data. UGPA = end-of-college GPA; HSGPA = high school cumulative GPA; XC.Y(U) = correlation between SAT scores and Immersion–Emersion racial identity schema (culture) with end-of-college GPA controlled; XC.Y(H) = correlation between SAT scores and Immersion–Emersion racial identity schema with high school GPA controlled; SAT-V = SAT-Verbal; SAT-M = SAT-Mathematics. Immersion–Emersion scores are T scores ($M = 50$, $SD = 10$) based on Bradby et al.'s (2001) sample.

- tive ability testing? *American Psychologist*, 47, 1083-1101.
- Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement*, 8, 1-4.
- Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education. *American Psychologist*, 56, 302-318.
- Wilson, K. M. (1981). Analyzing the long-term consequences of minority and nonminority students: A tale of two studies. *Research in Higher Education*, 15, 351-375.

Correspondence concerning this comment should be addressed to Janet E. Helms, Boston College, Department of Counseling, Developmental, and Educational Psychology, 317 Campion Hall, Chestnut Hill, MA 02467. E-mail: helmsja@bc.edu