School Psychology Review
1984, Vol. 13, No. 1

CULTURAL BIAS IN WISC SUBTEST ITEMS: A RESPONSE TO JUDGE GRADY'S SUGGESTION IN RELATION TO THE *PASE* CASE

Tong-He Koh Aurelius Abbatiello
Bureau of Child Study, Chicago Board of Education

Caven S. McLoughlin
Kent State University

ABSTRACT

This study has sought to determine empirically whether seven items from the Information and Comprehension subtests of the Wechsler Intelligence Scale for Children discriminate against any racial groups. These items were singled out by Judge John F. Grady in his opinion in the *PASE* (*P*arents in *A*ction in *S*pecial *E*ducation) case, as being culturally biased against black children. A stratified random sample (N=360) of test protocols of Chicago public school children who were referred for a psychological evaluation were analyzed quantitatively and qualitatively. These children were part of the sample being considered by the judge. Main comparisons of percentage passing items for race, sex, and age groups showed no significant differences. Error analyses showed no significant "cultural" differences between white and black children, in that none of the responses that were said to be likely to occur from blacks were evident.

In July 1980 United States District Judge John F. Grady of the Northern District of Illinois, Eastern Division, ruled that intelligence tests used in the Chicago public schools to diagnose children as mentally retarded are *not* biased against blacks. This decision was related to a suit initiated in 1974 by *P*arents in *A*ction in *S*pecial *E*ducation (*PASE*) against the Chicago Board of Education and the Illinois State Board of Education. The plaintiffs filed on behalf of two girls who had allegedly been misdiagnosed as retarded and placed in special education classes, who claimed that the misassessment was a direct result of inherent racial bias in the administered standardized intelligence tests.

Judge Grady personally examined all questions posed by three intelligence tests: Stanford-Binet, Form L-M (S-B); Wechsler Intelligence Scale for Children (WISC); and Wechsler Intelligence Scale for Children-Revised (WISC-R). He concluded that one item of the S-B (aesthetic comparison) and eight items of the WISC and WISC-R tests may possibly have been so culturally biased against black children, or at least sufficiently suspect, that their use was inappropriate. He recorded, however, that his view was not based on empirical findings. In his 117-page opinion Judge Grady noted that "despite the prodigious volume of test papers which has been accumulated over the past half century, there has been no extensive study undertaken to determine in specific terms just how blacks and whites compare to each other on all test items" (p. 102). Furthermore, he stated, "it would have been helpful to the court if plaintiffs had produced the actual scoring sheets which would have shown the verbatim responses of children. The production of that kind of evidence would have been far preferable to these almost casual recollections of (witnesses), the accuracy of which has to be taken on blind faith" (p. 44).

Literature reviews produced few studies that compared the performance of white and black children on individual test items. Mercer and Brown (1973) charged that the Comprehension test of the WISC was the most blatant manifestation of Anglo-centrism, since it requires not only a knowledge of Anglo value systems, but agreement with that set of values. In the CBS television documentary, "The IQ Myth" (1975), Dr. Robert Williams demonstrated the WISC Comprehension question: "What is the thing to do if a boy/girl much smaller than yourself starts to fight with you?" as one example of cultural bias. The correct answer, according to the manual, is one which demonstrates restraint such as calling the teacher or walking away. Thus, the black child whose value system dictates an appropriate response of striking back is penalized. In the same program David Wechsler, the

Address correspondence and requests for reprints to Tong-He Koh, 6007 N. Sheridan Road, 10B, Chicago, IL 60660.

WISC author, stated that a ghetto child should, in fact, receive credit for such a response.

Miele (1979) examined the rank order difficulty of this "fight" question across ethnic groups. Of the 161 WISC items, it ranked 42 for the black group and 47 for the white group. Thus, it may be deduced that this particular item was relatively easier for black than for white children.

In an analysis of incorrect responses to the Peabody Picture Vocabulary Test (PPVT), Jensen (1976) found that the errors for each item were distributed in a non-chance fashion over the multiple choice distractors in the same proportions for whites and blacks. Jensen (1974) examined the item difficulty levels of the PPVT and Raven Progressive Matrices for white, black, and Mexican-American ethnic groups in a California school district and reported that rank order correlations ranged from .86 to .99. Berry (1977) compared the rank order of PPVT item difficulties in white and black children in Middletown, Connecticut and found "no statistical significant difference in the correlation between item order and item difficulty for groups of different race or sex" (p. 40).

Eells (1946) compared the performance of "ethnic" (at least one foreign-born parent) and "old American" (both parents American-born) groups on a total of 650 single test items from the Otis-Lennon Mental Ability Test, Henmon-Nelson Tests of Mental Ability, Thurstone Primary Mental Abilities, and the California Test of Mental Maturity. The item analysis did not reveal any appreciable item index differences between the two groups. McGurk (1951) compared black and white performance on the Otis-Lennon Mental Ability Test, Thorndike's CAVD (Completions, Arithmetical problems, Vocabulary, and Directions) test and the ACE (American Council on Education) Psychological Examination for College Freshmen, between those items classified as the most cultural and those as the least cultural by a panel of 78 judges. They found that there was no significant difference between the two groups on the total test scores, and that blacks' performance was superior on those items judged as the most culture loaded.

Only two studies item analyzed the Stanford-Binet test for ethnic differences. Nichols (1972) found a .99 rank order correlation between the percentages of white and black children passing on 16 consecutive test items from age III-6 through V. Kennedy, Van de Riet, and White (1963) reviewed S-B protocols for 1,800 black school children in southeastern states, and compared them with the 1937 white standardization sample. The rank order correlations between the difficulty levels of the initial 26 words of the S-B Vocabulary test for the two groups was .98. These investigators also indicated that the percentage of the black sample passing each S-B item was a function of the item's mental age placement as determined in the white standardization sample.

Jensen (1977) examined the item difficulty of the Wonderlic Personnel Test collected from two independent samples of white and black job applicants. The cross-racial correlation between item difficulties was .93 and .96, respectively. In the area of scholastic achievement tests, Arneklev (1975) gave the five subtests of the Comprehension Tests of Basic Skills (Form Q, Level 3), published by the California Test Bureau, to school children in Tacoma, Washington, and found that the mean white-black difference in item difficulties, as measured by percent passing, was not significant.

Cardall and Coffman (1965) investigated item bias in the Scholastic Aptitude Test by means of analysis of variance and correlation of item difficulties, separately for the Verbal and Math subtests, across groups from three regions. The cross-racial correlation of item difficulties for Math was .89, and .84 for Verbal items. Angoff and Ford (1973), using the same item correlation method on the Preliminary Scholastic Aptitude Test, correlated item difficulties across randomly selected groups of white and black groups from a large urban area in the Southwest. The correlations of item difficulties between the randomly selected white and black groups ranged from .86 to .95. Unfortunately, none of these investigations included a mentally retarded population. It is with the population of black and white mentally retarded children that the *PASE* litigation is concerned. This study was designed to test Judge Grady's assertion that seven WISC items specifically discriminate against black children due to inherent cultural bias.

There are many ways to conceptualize and analyze test or item bias. Jensen (1979)

describes both internal and external classes of procedures. Internal procedures (e.g., rank order of item difficulty and factor analyses) examine how children perform on the item itself (Miele, 1979; Sandoval, 1979). External procedures, (e.g., regression analysis) are based on how well items predict future performance (Reschly & Sabers, 1979; Reynolds & Hartlage, 1979). Following the model of construct validity, Shepard (1982) proposed two techniques: logical and empirical analysis. Logical analysis is accomplished through a deduction from theory to external observables. In construct validation, logical analysis or conceptual analysis remains the only means for formally elaborating what is to be measured and how this trait is expected to relate to other variables.

Numerous empirical methods have been proposed for operationalizing item bias. These include differences in item difficulty (Angoff & Ford, 1973), group differences in item discriminations (Green & Draper, 1972), comparisons of item characteristic curve (Durovic, 1975; Lord, 1977; Wright, Mead, & Draba, 1976), and distractor or error analysis (Scheuneman, 1982). A further method employed to identify bias involves a panel of judges and is sometimes referred to as face valid or "armchair analysis." Williams (1971), for example, pointed out that certain items of the WISC Comprehension subtest (e.g., item 6) may be more difficult for black children to answer because of differences in culture and experience. Reynolds (1982) summarized several studies that have found dismal evidence for the effectiveness of such methods in finding biased test items (Jensen, 1977; Plake, 1980; Sandoval & Miille, 1980). When asked to subjectively review items, judges were consistently unable to identify items that were relatively more difficult for black or Chicano than for white children. Even so, Helmstadter (1964) and Anastasi (1976) assert that face validity has a place in testing, specifically in gaining rapport and maintaining good public relations.

Although not as widely used as the WISC-R, there are still a substantial number of WISC administrations. However, the significance of this study is not so much one of providing evidence for lack of bias in a once commonly used test, as one of testing the armchair hypotheses of the judge in litigation of critical relevance to school psychology. The primary purpose of this study was to contrast two methods of item bias identification: judgmental and empirical. The armchair judgment of item bias was based on the opinion of Judge Grady, whereas the empirical judgment was determined by an item analysis of black and white educably mentally retarded children's performance on seven cited WISC subtest items.

Method

This study analyzed WISC protocols of children with full scale IQ scores from 55 to 85, who were not identified as either physically or emotionally handicapped, tested by the Bureau of Child Study, Chicago Board of Education, from 1975 to 1977. From this pool of over 30,000 protocols a stratified random sample of 360 protocols was selected equal in number for each of the stratification variables of race (black and white), sex (males and females), and age (primary, intermediate, advanced), as shown in Table 1. The white children had a mean age of 11.27 years (SD=26.93 mo., range: 7.00-15.16 yrs.); and blacks of 11.30 years (SD=22.85 mo., range: 7.00-15.67 yrs.); with mean FSIQs for whites and blacks of 74.63 (SD=7.78, range: 55-85) and 75.52 (SD=7.00, range: 56-85), respectively.

The seven WISC items cited in the legal opinion as being potentially culturally biased against black children (three Information test items: Rubies, Cash on Delivery, and Stomach; and four Comprehension items: Fights, Loaf of Bread, Pay Bills by Check, and Give Money to Charity) were analyzed both quantitatively and qualitatively. Quantitative analyses were computed through a consideration of pass-fail, and qualitatively by an analysis of the content of responses for each protocol item.

Results

Results are presented in two sections to reflect quantitative and qualitative interpretations: percent passing and error analysis.

Quantitative Analysis

The percentage of children passing each of the seven items was determined for each of the 12 race × sex × age groups. For the three Information items, responses were scored 1 for those passing and 0 for those failing. The four Comprehension items were scored 2, 1 (pass) or 0 (fail). For purposes of analysis, scores of 1 and 2 were combined as "plus." All scoring followed Manual (Wechsler, 1949) procedures.

The percent passing data for the seven WISC items according to the 12 race × sex × age groups is presented in Table 2. The percent passing for the combined groups with calculations of *Chi-square* values for the main comparisons of race, sex, and age and subgroups are displayed in Table 3.

No significant differences were found between the black and white groups on any of the seven contentious items. A significant sex difference was found for only one item; Rubies. As a group, more girls knew the color of rubies than boys (χ^2=6.40, df=1, p<.01). Further, it was noted that white girls scored significantly higher than black girls on this item (χ^2=4.06, df=1, p<.05). Further analyses revealed that this phenomenon occurred within the advanced age subgroup; that is, older white girls knew better the functions of the stomach than the older black girls (χ^2=5.46, df=1, p<.02).

TABLE 1

Means and standard deviations of chronological age (CA) and WISC full scale IQ (FSIQ) by race × sex × age groups

	White Males	White Females	Black Males	Black Females
Primary Age				
7-9 yrs.	N=30	N=30	N=30	N=30
CA *m*	103.50	103.90	102.80	103.07
sd	10.19	8.17	9.59	9.43
FSIQ *m*	76.43	74.93	75.43	74.23
sd	6.69	7.44	7.49	6.54
Intermediate Age				
10-12 yrs.	N=30	N=30	N=30	N=30
CA *m*	137.53	136.47	137.90	136.87
sd	10.47	10.65	10.23	11.30
FSIQ *m*	74.87	74.30	75.47	75.70
sd	6.08	7.59	7.21	7.43
Advanced Age				
13-15 yrs.	N=30	N=30	N=30	N=30
CA *m*	166.50	165.27	166.80	167.83
sd	7.43	6.68	6.76	7.70
FSIQ *m*	74.87	72.37	74.10	71.57
sd	8.59	9.92	6.29	6.78

TABLE 2

Percent of children passing each of the seven WISC items in the twelve race × sex × age groups

	Group Membership[1]											
Item	WMP	WMI	WMA	WFP	WFI	WFA	BMP	BMI	BMA	BFP	BFI	BFA
Rubies	13	30	37	27	43	60	13	23	30	20	30	37
Stomach	0	3	20	0	13	30*	0	3	27	0	7	7*
C.O.D.	0	0	0	0	0	0	0	0	0	0	0	7
Loaf of Bread	60	73	83	37	77	73	60	77	93	53	77	83
Fight	47	80	83	60	77	80	47	83	97	43	87	83
Pay bills by check	3	20	43	3	13	47	0	17	43	0	30	37
Give money to charity	0	10	20	0	13	30	3	7	33	0	13	13

[1]W or B in the first position refers to white or black; M or F in the second position refers to male or female; and P, I, or A in the third position refers to primary, intermediate, or advanced age groups. Each combination group is N=30.
*χ^2×5.46. df=1, p<.02.

TABLE 3

Percent of children passing each of the seven WISC items in combined sex × race groups

	Group Membership							
Item	White Boys N=90	White Girls N=90	Black Boys N=90	Black Girls N=90	Total Boys N=180	Total Girls N=180	Total Whites N=180	Total Blacks N=180
Rubies	27	43*	22	29*	24*	36*	35	26
Stomach	8	14*	10	4*	9	9	11	7
C.O.D.	0	0	0	2	9	1	0	1
Loaf of Bread	72	62	77	71	74	67	67	74
Fight	70	72	76	71	73	72	71	73
Pay bills by check	22	21	20	22	21	22	22	21
Give money to charity	10	14	14	9	12	12	12	12

*χ^2, df=1, p<.05.

Qualitative Analysis

Incorrect verbatim responses recorded on the protocols were analyzed to test the hypothesis that white and black children make qualitatively different errors based on their respective cultural perceptions.

1. *Rubies:*
   During the trial, Dr. Robert Williams, an expert witness summoned by attorneys for the plaintiffs, criticized this item as confusing to black children because "Ruby" can be used as a woman's name. Dr. Williams testified that he had heard a little boy say, "Well, she is black." Incorrect responses given on the sample of test record forms were primarily the names of 14 different colors, including black and white. There was no single response given by either a white or black child who thought "ruby" was a girl's name.
2. *Stomach:*
   Dr. Williams testified that many black children answer, "It growls." He attributed this to an assertion that many black children come from poverty-level families where they have insufficient nourishment. The number of "growl" responses by white and black children was 29 (39%) and 39 (48%) respectively, of the total sample of incorrect responses. This difference between the two groups was not statistically significant (χ^2=.24, df=1, p>.50).
3. *C.O.D.:*
   Dr. Williams testified that a significant number of black children have had insufficient exposure to the term "C.O.D." to be able to appropriately respond. However, he contended, this does not mean that black children are unable to codify abbreviations and symbolize these abbreviations at a higher cognitive level.

   The "C.O.D." item appears as number 18 among 30 items. Only two children of the total sample of 360 gave a correct response to this item, and these children were black girls from the advanced age group. Thus, it may be stated that most children in this study did not reach this level of difficulty, and are unlikely to have known the meaning of "C.O.D."
4. *Loaf of Bread:*
   Dr. Williams again indicated, "the correct response is really culturally determined, because kids used to tell me 'well, I go back home because my mama told me don't be foolin' around on the street, that if I go to the store, don't get lost, don't go any other place, because I'm going to beat you!' " In the investigated sample, 47 (20%) white children responded in a way that indicated "go back home and tell Mom," and 38 (17%) blacks gave an equivalent response.
5. *Fight:*
   This "fight" question is perhaps the most famous "cultural" item in the intelligence test controversy and is most frequently cited by critics of intelligence tests as being an example of serious bias. The reason for the assertion of bias, according to Williams and some other critics, is that in black communities children are purposefully taught that if anyone hits them they should retaliate by hitting back. The previously cited Miele (1979) study was the only formal empirical investigation of the possibility of cultural bias, which determined that this question was advantageous to black children. In the present sample, the findings were in accord with Miele (1979), for 34 (71%) of white children and 28 (61%) of the black children gave a response of "fight back" or similar answers such as "punch him," "kick her," and "knock him down."
6. *Pay Bills by Check:*
   This item is subjected to the same criticism as the "C.O.D." item which relates to an assertion of relative inexperience for many black children. The most common incorrect responses for both ethnic groups in this study came from a confusion between check and cash; examples include: "to save money," "don't have enough money," "need money for other things." There were no patterns of errors to suggest a different history of experience between the two groups.
7. *Give Money to Charity:*
   Dr. Williams asserted that an economically poor child, or a child on welfare, would be less likely to donate money to an organized charity than to a blind or crippled person he saw on the street. One of the zero responses recorded in the manual for this item is, "if you give it to a beggar, he is liable to keep it to himself." Williams suggested that this is exactly what one would want the beggar to do and that such a response is therefore appropriate. The data of this study revealed that eight white and three black children thought that the beggar should keep the money. Black children gave more qualitatively negative comments regarding beggars than did white children, saying for example, that they "can't be trusted," "buy booze," and "they are no good."

In summary, the error analysis of the seven contentious, purportedly culturally biased items, showed no significant differences between white and black children. None of the responses that were said to be likely to occur in blacks because of alleged cultural and experiential biases were evident in this analysis.

Discussion

This study has sought to empirically determine whether seven items from the Information and Comprehension subtests of the WISC discriminate unfairly against black children. These items were singled out by Judge John F. Grady, in a landmark legal decision relating to the use of intelligence measures for children, as being either culturally biased or at least to be sufficiently suspect that they should not be considered when making special education placement decisions. The results demonstrate that no such condition exists. There were only two isolated instances, in one group, in which statistically significant differences did occur. Thus, this evidence offers strong support to Judge Grady's contention that the possibility for these anomalies to cause an "educable mentally handicapped" (EMH) placement, that would not otherwise occur, is practically non-existent.

The authors recognize the limited generalizability of this conclusion which deals with a very specific sample of EMH children from the Chicago public school system. Nevertheless, it was these very children upon whom the *PASE* decision was focused. The assertion of the plaintiffs that the children who constituted this sample could be discriminated on the basis of ethnicity, is not supported by the data. It appears that the experience and expo-

sure that all children, whether white or black, receive in a large urban city like Chicago are sufficiently similar that WISC items are not biased in favor of any one ethnic group.

Further analysis of all the items of the Wechsler intelligence tests for children would seem to be warranted to demonstrate empirically whether there is any inherent bias. Such a study would put to rest any doubts reflected in "armchair" inspection of the intelligence test items. □

REFERENCES

Anastasi, A. *Psychological testing* (4e). New York: Macmillan, 1976.

Angoff, W. H., & Ford, F. F. Inter-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, 1973, *10*, 95-106.

Arneklev, B. L. *Data related to the question of bias in standardized testing.* Tacoma, WA: Washington Office of Education, Tacoma Public Schools, 1975.

Berry, G., Jr. *An investigation of the item ordering of the Peabody Picture Vocabulary Test by sex and race.* Doctoral dissertation, University of Connecticut, 1977.

Cardall, C., & Coffman, W. E. *A method for comparing the performance of different groups on the items in a test.* Research Bulletin: 64-61. Princeton, NJ: Educational Testing Service, 1964.

Columbia Broadcasting System. *The IQ myth. Special Report.* April 22, 1975.

Durovic, J. J. *Definitions of test bias: A taxonomy and an illustration of an alternative model.* Unpublished doctoral dissertation, State University of New York at Albany, 1975.

Eells, K. *Intelligence and cultural differences.* Chicago, IL: University of Chicago Press, 1946.

Grady, J. F. *Legal opinions on the suit: Parents in Action on Special Education et al. v. Joseph P. Hannon et al.* United States District Court for the Northern District of Illinois, July, 1980.

Green, D. R., & Draper, J. F. *Exploratory studies of bias in achievement tests.* Paper presented at the annual meeting of the American Psychological Association, Honolulu, 1972. (ERIC Document Reproduction Service No. ED 070 794).

Helmstadter, G. C. *Principles of psychological measurement.* New York: Appleton-Century-Crofts, 1964.

Jensen, A. R. How biased are culture-loaded tests? *Genetic Psychology Monograph*, 1974, *90*, 185-244.

Jensen, A. R. Test bias and construct validity. *Phi Delta Kappan*, 1976, 340-346.

Jensen, A. R. An examination of culture bias in the Wonderlic Personnel Test. *Intelligence*, 1977, *1*, 51-64.

Jensen, A. R. *Bias in mental testing.* New York: Free Press, 1979.

Kennedy, W. A., Van de Riet, V., & White, J. C. A normative sample of intelligence and achievement of Negro elementary school children in the southeastern United States. *Monographs of the Society for Research on Child Development*, 1963, *28*, No. 6.

Lord, F. M. A study of item bias using item characteristic curves theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology.* Amsterdam: Swets and Zeitlinger, 1977.

McGurk, F. C. J. *Comparison of the performance of Negro and white high school seniors on cultural and noncultural psychological test questions.* Washington, DC: Catholic University Press, (Microcard), 1951.

Mercer, J. R., & Brown, W. C. Racial differences in IQ: Fact or artifact? In C. Senna (Ed.), *The Fallacy of IQ.* New York: Third World Press, 1973, 56-113.

Miele, F. Cultural bias in the WISC. *Intelligence*, 1979, *3*, 149-164.

Nichols, P. L. *The effects of heredity and environment on intelligence test performance in four- and seven-year-old white and Negro sibling pairs.* Unpublished doctoral dissertation, University of Minnesota, 1972.

Plake, B. S. A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. *Educational and Psychological Measurement*, 1980, *40*, 397-404.

Reynolds, C. R. The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology.* New York: Wiley, 1982.

Reynolds, C. R., & Hartlage, L. Comparison of WISC and WISC-R regression line for academic prediction with black and with white referred children. *Journal of Consulting and Clinical Psychology*, 1979, *47*, 589-591.

Reschly, D., & Sabers, D. Analysis of test bias in four groups with regression definition. *Journal of Educational Measurement*, 1979, *16*, 1-9.

Sandoval, J. The WISC-R and internal evidence of test bias with minority groups. *Journal of Consulting and Clinical Psychology*, 1979, *47*, 919-927.

Sandoval, J., & Miille, M. P. W. Accuracy of judgments of WISC-R item difficulty for minority groups. *Journal of Consulting and Clinical Psychology*, 1980, *48*, 249-253.

Scheuneman, J. D. A posteriori analyses of biased items. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias.* Baltimore: The Johns Hopkins University Press, 1982.

Shepard, L. A. Definition of bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias.* Baltimore: THe Johns Hopkins University Press, 1982.

Williams, R. L. Abuses and misuses in testing black children. *Counseling Psychologist*, 1971, *2*, 62-73.

Wright, B. D., Mead, R. J., & Draba, R. *Detecting and correcting test item bias with a logistic response model* (Research Memorandum No. 22). Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1976.