A Cross-Cultural Analysis of the Fairness of the Cattell Culture Fair Intelligence Test Using the Rasch Model

H. Johnson Nenty
Cleveland Public Schools

Thomas E. Dinero
Kent State University

Logistic models can be used to estimate item parameters of a unifactor test that are free of the examinee groups used. The Rasch model was used to identify items in the Cattell Culture Fair Intelligence Test that did not conform to this model for a group of Nigerian high school students and for a group of American students, groups believed to be different with respect to race, culture, and type of schooling. For both groups a factor analysis yielded a single factor accounting for 90% of the test's variance. Although all items conformed to the Rasch model for both groups, 13 of the 46 items had significant between score group fit in either the American or the Nigerian sample or both. These were removed from further analyses. Bias was defined as a difference in the estimation of item difficulties. There were six items biased in "favor" of the American group and five in "favor" of the Nigerian group; the remaining 22 items were not identified as biased. The American group appeared to perform better on classification of geometric forms, while the Nigerians did better on progressive matrices. It was suggested that the replicability of these findings be tested, especially across other types of stimuli.

Studies of heredity, environment, and intelligence since the early 1930s have shown that environmental conditions do significantly influence intelligence test scores (Hunt, 1972; Jordan, 1933; Leahy, 1935; Neff, 1938; Wellman, 1934, 1937). Any test is assumed to be a measure of a "phenotype" (Jensen, 1980) and, indeed, must be influenced by the environment to some extent. Davis (1948), Davis and Havighurst (1948), and Eells, Davis, Havighurst, Herrick, and Tyler (1951), however, have maintained that all intelligence tests fail to measure general ability directly because they discriminate between social classes, cultures, and school and home training and are influenced by motivation, emotional reaction, test sophistication, and speed. Wellman (1937), and later Hunt (1972), provided evidence to show that improvements in the early childhood environment, such as attendance at preschool clinic, are influential in raising and possibly maintaining a child's IQ. As a result it could be concluded that there exists a large set of factors on which IQ scores depend and from which a child's IQ could be fairly predicted without the child ever being tested. Many of these authors also imply that many of these characteristics are inextricably confounded with intelligence so that, if indeed, there are valid group differences in intelligence, these differences would never be measured directly.

Prior to Cattell's (1940) work, attempts to measure intelligence appear to have been based mostly on Spearman's single factor theory of general ability. Cattell, on the other hand, concluded that second-order analyses or primary abilities result in two main factors (see also Cat-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 5, No. 3, Summer 1981, pp. 355–368

0146-6216/81/030355-14$1.70

tell, 1968, p. 56): one referred to as *fluid* and the other *crystallized* intelligence factors. Fluid intelligence is said to be the "major measurable outcome of the influence of biological factors on intellectual development—that is heredity" (Horn & Cattell, 1966, p. 254); and crystallized intelligence can be accounted for by the investment of fluid intelligence on environmental facilities such as early exposure to intellectually stimulating experiences, school, and so forth (Cattell, 1979, p. 5). Crystallized general intelligence shows itself in skills that have been acquired by cultural experiences, such skills that traditional intelligence tests in part are seen to measure. By differentiating between two essentially different factors of general ability, Cattell not only provided a parsimonious model of an obscure construct, intelligence, but also provided a basis for a more meaningful measure of that construct. From this, it seems that crystallized intelligence should be heavily dependent on culture and fluid intelligence should not; and even though differences might exist within a culture among individuals' fluid intelligence, there are not necessarily any differences among cultures.

Recognizing fluid intelligence as a distinct factor, Cattell and his colleagues began to search for a more "saturated" measure to define it operationally. The results turned out "almost as a by-product, to have promising properties as culture-fair tests" (Cattell, 1971, p. 82). Since then, "culture-fair intelligence tests . . . have become the practical test of expression of fluid general ability factor" (Cattell, 1971, p. 78). Though culture-fair tests are found to measure mostly fluid intelligence, the converse is not true; that is, measures of fluid intelligence are found only among noncultural performances (Cattell, 1979, p. 6). Measures of fluid intelligence include tests of judgments and reasoning involving classification, analogies, matrices, and topology (Cattell, 1968, p. 58); and these are, indeed, the types of stimuli that Jensen (1980) refers to as "culturally reduced." Cattell's culture-fair test was designed to "avoid relationships depending on anything more than immediately given fundaments that are equally familiar or equally unfamiliar to all" (Cattell & Horn, 1978, p. 141).

The Cattell Culture Fair Intelligence Test (CCFIT) is assumed to be indifferent to cultural experiences that might differentially influence examinees' responses to its items (Cattell, 1968, p. 61) and also is assumed to be saturated with only one single second-order intelligence factor—fluid intelligence (Cattell, 1971, p. 78). Given these assumptions, it could be said that it is a unifactor or unidimensional measure of fluid intelligence, which (according to Cattell's theory) does not depend on cultural experiences.

Though many experts in cross-cultural psychology generally deny the possibility of a culture-free test (Dague, 1972; Manaster & Havighurst, 1972), it is believed that with modern measurement techniques, tests could be designed and administered to *minimize* the influence of cultural experiences and to maximize the influence of the trait that the test is designed to measure.

Measurement models have been proposed whereby objective measurement of a single intrinsic factor (Biesheuvel, 1952) can be achieved. The logistic model attributed to Rasch (1960) specifies a probabilistic relationship between the observable item and person scores and the unobservable item and ability parameters assumed to underlie the result of a person-by-item interaction in a testing situation. Specifically, the model holds that, given a unidimensional test, the probability of any examinee succeeding on any item is wholly determined by his/her ability on the trait being measured and by the difficulty parameter of the item. For this model "the estimate of the item difficulty parameter will not vary significantly over different samples . . . [but the] item parameter will not be sample-invariant when there is culture bias which differentially affects the item probability and alters the [item characteristic curve] slopes" (Whitely & Dawis, 1974, p. 175). This implies that "if a test includes items which differ in cultural loadings, the special conditions required for item parameter invariance may be difficult to obtain"

(Whitely & Dawis, 1974, p. 176). Some cross-cultural psychologists have strongly recommended the use of this model in refining psychological tests (Eckensberger, 1972, p. 103; Fisher & Pendl, 1980; Massad, 1978). To the extent that a test fails to meet the theoretical requirements of culture-fairness, data from such a test will not fit the Rasch model. Indeed, Mellenbergh (1972) was not able to fit the model to data from Dutch and Indonesian junior high students. He suggested that local influences, such as whether or not specific material is taught, can randomly influence data.

Latent trait models, such as the Rasch, are indeed well formed to study cultural bias. Since the estimation procedures separately estimate item difficulties and person abilities, items can be selected that have difficulties estimated to be the same across groups (and therefore to be unbiased; see Ironson, 1978, for example), and then these items can be used to study individual or group ability differences.

The problem of the present study was stimulated by the fact that "developing countries show an increasing need for tests adapted to their own cultural conditions and requirements. As a common rule Western tests are used as the starting point but require modifications" (Drenth & Flier, 1976, p. 137), the extent of which would depend on the cultural differences. In Western cultures it has been observed that "testing has contributed to more effective use of manpower, to more equal distribution of educational and professional opportunities, and to identification of talents that might otherwise remain hidden" (Drenth, 1977, p. 23). It is generally accepted that "tests should be utilized for the same purposes in the developing nations whose traditional cultures are evolving towards that in which those ability tests are developed" (Iwuji, 1978, p. 16). In most of these developing nations—Nigeria, for example—decisions have to be made for large heterogeneous populations, and if the test scores on which such decisions are based are observed from tests that are culturally biased, then such decisions are not impartial.

Most Western ability tests are culture-dependent measures of crystallized intelligence. Using these in developing nations would not correct for the bias due to differences in the cultures, quality of schools, social status, sex, and other environmental factors. There were, then, two research questions for the present study. The first was concerned with how well each item of the CCFIT defines the construct of fluid intelligence for each of two disparate cultures, and secondarily with whether the total test is measuring the same construct for both of the cultures. In Rasch terminology, this concern translates to showing how well each CCFIT item's data fit the Rasch model for each of the two groups and then both samples combined: The extent to which an item's data fit the Rasch model indicates the accuracy with which that item defines that single trait which the model assumes the test is measuring.

The second research hypothesis was concerned with item bias. Rasch item parameters, as with all latent trait models, are assumed to be sample invariant, and for a given test the estimates of this parameter for different samples of examinees should not be significantly different from each other; but the "item parameter will not be sample-invariant when there is cultural bias which differentially affects the item probabilities" for the different sample of examinees (Whitely & Dawis, 1974, p. 175).

Within the Rasch model, the hypothesis of equal item parameter estimates could also be stated with regard to each item's characteristic curves derived from the data from each of the two groups. This is because, unlike other latent trait models, the Rasch model assigns only one parameter, item difficulty, to the item; that is, only item difficulties determine differences among item characteristic curves.

Method

The primary aim of the study was to cross-validate the CCFIT with two mutually remote racial groups that are, in addition, culturally disparate. The first sample was made up of 600 American junior high school students from four

Portage County, Ohio, schools. The second sample was made up of 803 Nigerian students from seven secondary schools in Lagos and Cross River States.

The two samples, composed mostly 13- to 15-year-old boys and girls, were comparable with respect to the number of years they had spent in formal schooling. Subjects were, by necessity, volunteers, but random sampling was not seen to be necessary because the Rasch model is sample free in its estimation of item parameters.

Culture fairness in testing cannot be achieved through test construction only but through a combination of test construction, test instruction, and test administration. Whereas test construction for culture fairness would attempt to equalize the influences of extraneous cognitive factors—such as language, test content, and materials—test instruction and administration for culture fairness would try to equalize, as much as possible, the influence of all noncognitive factors such as speed, test sophistication, motivation, rapport with the examiners, understanding of instructions, and understanding of what one is expected to do. The main aim was to rule out, or to hold constant, parameters on which cultures might differ, except for the trait of interest. Through construction the CCFIT purports to rule out or hold constant the influence of contaminating cognitive factors. In this study, then, an attempt was made to control additionally for possible noncognitive extraneous factors. Instructions and directions supplementary to the standardized set for the administration of the test were given to each examiner. These allowed for a motivation session, a practice session for each subtest, and other specific instructions to the examiner on the conduct of the testing. The normal time allowed for the testing was also doubled.

The CCFIT is a paper-and-pencil perceptual test that consists of abstract geometric forms. It consists of four subtests selected because they correlate well with Spearman's general mental capacity. In the first, the task of the examinee is to select from five choices the one that best completes a progressive series. In the second subtest the examinee is required to identify from among five figures the one that is different from the others. The third subtest requires the examinee to complete a matrix of geometric figures by choosing from one of five figures presented. In the last subtest the examinee's task is to select from five choices the one that duplicates a topological relationship among geometric forms that is the same as the one given.

Scale 2 Form A of the test was used. The test is reported to have a Spearman-Brown reliability coefficient of .79, a Cronbach's alpha of .77, and a KR21 of .81 (Institute of Personality and Ability Testing, 1973, p. 10).

The Rasch model is based on very stringent theoretical assumptions, the consequences of which include such requirements as absolute independence of examinees' responses to each item, maximum level of motivation by the examinee so as to apply all his/her related ability in answering each item of the test, no time constraints in the testing, equality of item discriminations, and the absence of guessing. In other words, the theory considers item difficulty and person ability as the only dynamics that underlie the result of any person-by-item interaction. Such requirements could rarely be met in practice; thus, working on the assumptions that they had been met constitutes a limitation to the study. Some of the changes made on the normal instructions for the administration of the CCFIT were done in an attempt to make the testing conditions compatible with the requirements of the Rasch model and culture fairness in test instruction and administration.

Latent trait theory assumes unidimensionality of the test space, and since Stelzl (1979) has indicated that violation of this assumption can still result in artificial fit to the data, a first screening of the item data's fit to a Rasch model was a factor analysis. Assuming that the trait of intelligence is normally distributed and that phi coefficients would lead to spurious results (Lord & Novick, 1968), the matrix of tetrachoric correlations among items was calculated and used as input in a principal axis factor analysis. Then, based on a rationale proposed by Wright (1977a) and also by Wright, Mead, and Draba (1976), Rasch test and item calibrations were performed for each of the cultural groups, the two groups combined, and a random division of the combined sample.

Estimating the parameters of the model can be done using either a procedure where the person's abilities are estimated and the test's item difficulties are estimated conditional on these (Anderson, 1972) or by reciprocally estimating one set, then the other, until adjustments yield the desired precision. While Anderson had indicated that this latter (unconditional) procedure leads to inconsistent estimates, Wright and Douglas (1977) have suggested that the procedure is impractical for calibrating more than, say, 15 items; and, indeed, if the mean item difficulty is zero, the two procedures yield no discernible differences. The present data conformed to this criterion; hence, the present calibrations were performed using the unconditional maximum likelihood procedure as devised by Wright and Panchapakesan (1969), developed and used by Wright and Mead (1978) in their computer program BICAL.

Results

Selected subgroups' total test raw score means and standard deviations are presented in Table 1. For both cultural groups, there appears to be minimal variation across subgroups. For each group the factor analysis yielded one factor with an eigenvalue of 41.4, accounting for 90% of the item pool's variance in the American sample and an eigenvalue of 41.6 for the Nigerian group, also accounting for 90% of the variance. It was thus assumed reasonable to proceed with the item calibration.

Table 1
Descriptive Statistics for Both Groups' CCFIT Scores by Sex, Age, and Grade

Cultural Group		Sub-group	N	Mean	SD
Portage County (U.S.A.)	Sex:	Male	284	30.93	6.33
		Female	316	31.36	6.03
	Age:	<13 years	73	30.48	6.24
		13 years	297	30.78	6.02
		14 years	209	32.32	6.04
		15 years	18	27.94	8.00
		>15 years	3	28.01	7.97
	Grade:	7th	295	30.10	6.19
		8th	305	32.22	5.94
	Total		600	31.18	6.18
Nigeria	Sex:	Male	614	30.30	5.97
		Female	189	29.63	5.18
	Age:	<13 years	75	31.92	4.96
		13 years	179	29.93	5.81
		14 years	232	30.07	6.33
		15 years	163	30.00	5.36
		>15 years	154	29.92	5.06
	Grade:	Class 2	443	30.00	5.89
		Class 3	360	30.33	5.09
	Total		803	30.15	5.68

Item Calibration

For the American sample, after the initial parameter estimation, seven persons with significant fit statistics ($t > 2.00$) were omitted from the final calibration in order to achieve more stable item parameter estimates for the group. For the Nigerian group, 10 such scores were observed and these persons were removed from final calibration. For each of the two groups, the item parameter estimates, along with selected fit statistics, and the discrimination index of each item in each of the four subtests are given in Tables 2 through 5. The item discriminations are assumed to be the same for all items within any test used to define a variable. Although in most Rasch analyses they are not discussed, they are presented here as an indication of how well individual items met this assumption.

Table 2
CCFIT Item Parameter Estimation for
Subtest 1 for Both Groups

Item	Item Difficulty	Between Group Fit t-test	Total Fit[a] Mean Square	Total Fit t-test	Disc. Index
Portage County (U.S.A.)					
1	-2.58	-.76	.78	-1.34	1.19
2	-2.64	.46	.86	-.79	1.37
3	-.95	-1.99	.94	-.65	.95
4	-1.78	1.92	.90	-.76	1.09
5	-.82	2.75*	1.01	.12	.58
6	-.74	1.22	.92	-.91	1.09
7	-.17	.82	.97	-.33	1.05
8	-1.36	1.60	.89	-1.08	1.36
9	-1.41	2.62*	.84	-1.54	1.48
10	2.00	-.90	1.01	.12	.97
11	1.18	-1.56	1.01	.19	.95
12	.47	-.53	1.02	.31	.87
Nigeria					
1	-2.81	-.62	.87	-.83	1.25
2	-2.37	-.60	.92	-.61	1.13
3	-1.33	.91	.05	-1.13	1.04
4	-2.07	.16	.89	-.95	1.26
5	-.93	1.02	.98	-.26	.88
6	-.43	.11	.97	-.46	1.01
7	.12	-1.43	1.00	.09	1.01
8	-.34	-1.74	.98	-.30	1.02
9	-.62	1.21	.92	-1.26	1.29
10	1.60	-.77	1.00	.06	.92
11	1.36	2.21	1.07	1.35	.62
12	.76	1.37	1.00	-.04	.99

*$p < .01$

[a]Weighted

Table 3
CCFIT Item Parameter Estimation for
Subtest 2 for Both Groups

Item	Difficulty	Between Group Fit t-test	Total Fit[a] Mean Square	Total Fit t-test	Disc. Index
Portage County (U.S.A.)					
1	-2.93	1.10	.90	-.44	1.04
2	-1.50	-.87	.94	-.53	.99
3	-2.58	.51	.91	-.49	1.34
4	.98	1.23	1.01	.10	1.02
5	1.03	2.95*	1.10	1.49	.67
6	-2.14	.28	.96	-.24	1.18
7	2.40	.58	.97	-.34	1.05
8	2.06	.76	1.01	.19	.97
9	-.07	.76	.10	.76	.94
10	2.36	3.41*	1.14	1.87	.49
11	.12	-2.28	1.03	.41	.96
12	1.99	2.78*	1.07	1.12	.61
13	1.97	1.35	1.02	.31	.89
14	3.67	4.27*	1.07	.70	.41
Nigeria					
1	-2.67	-.30	.92	-.49	1.32
2	-.94	-.29	.97	-.32	1.15
3	-1.36	.56	.96	-.43	1.20
4	.96	.16	1.04	.84	.81
5	1.07	3.99*	1.10	1.88	.45
6	-1.61	1.71	.96	-.43	1.34
7	2.26	.98	.98	-.35	1.11
8	2.41	1.84	1.01	.17	.81
9	.15	-.02	1.01	.22	.93
10	1.67	2.23	1.06	1.02	.66
11	.43	3.38*	1.08	1.39	.62
12	1.60	.42	.81	1.04	.75
13	1.92	1.41	.98	-.32	1.09
14	3.06	3.97*	1.09	1.16	.29

*$p < .01$

[a]Weighted

Test of Fit

Two important test-of-fit statistics are associated with each Rasch item parameter estimate. One is the total fit mean square with its total fit t statistic, and the other is the between-group fit statistic. The total fit mean square is the Rasch index of disagreement between the variable measured by a given item and the one that the model assumes. The expected value of this index

Table 4
CCFIT Item Parameter Estimation for
Subtest 3 for Both Groups

Item	Item Difficulty	Between Group Fit t-test	Total Fit[a] Mean Square	Total Fit t-test	Disc. Index
Portage County (U.S.A.)					
1	-2.47	.50	.87	-.79	1.11
2	-2.58	.24	.83	-.99	1.24
3	-2.77	.06	.87	-.65	1.34
4	-.69	.33	1.00	.33	.84
5	.31	1.04	.91	-1.30	1.30
6	-.70	3.49*	.83	-2.08	1.53
7	.55	-.99	.91	-.17	.99
8	.09	-2.19	.97	-.36	1.00
9	1.66	.97	1.05	.73	.83
10	.36	.58	.99	-.12	1.03
11	.22	.60	.94	-.93	1.18
12	.94	-1.80	1.00	-.02	1.04
Nigeria					
1	-2.33	-1.12	.91	-.70	1.10
2	-2.18	1.23	.97	-1.16	1.46
3	-2.27	.19	.88	-1.01	1.05
4	-1.45	.31	.91	-1.00	1.10
5	.33	2.64*	.93	-1.33	1.33
6	-1.10	3.48*	.87	-1.75	1.57
7	.37	-.00	.98	-.34	1.05
8	-.37	1.58	.93	-1.05	1.20
9	1.03	.11	1.00	.10	.93
10	.75	1.62	.96	-.82	1.22
11	-.00	2.02	.95	-.79	1.17
12	.83	2.01	1.02	.34	.93

*$p < .01$

[a]Weighted

is 1, and a significant total fit t statistic for an item signifies that there is a significant discrepancy between the data and the estimated model. On the other hand, a significant between-group fit statistic for an item indicates that statistically nonequivalent estimates of that item difficulty resulted from using different score groups from within the same sample (Wright & Mead, 1978, p. 77).

Based on the first fit criterion, no differences between the expected and the observed responses for each of the items were observed, at an alpha level of .01, for either cultural group, and for all the subjects combined. That is, all the CCFIT items fit the Rasch model for all three analyses. But for the two groups separately, a total of 13 of these "fitting" items had significant between-group fit statistics. That is to

Table 5
CCFIT Item Parameter Estimation for
Subtest 4 for Both Groups

Item	Item Difficulty	Between Group Fit t-test	Total Fit[a] Mean Square	Total Fit t-test	Disc. Index
Portage County (U.S.A.)					
1	-.10	.84	.96	-.49	1.19
2	.89	2.00	.94	-.99	1.24
3	1.21	1.40	.97	-.48	1.19
4	.24	-.24	1.00	-.04	1.04
5	.76	-.72	.98	-.31	1.10
6	.49	2.63*	.89	-1.75	1.43
7	.54	-1.57	.99	-.16	1.02
8	2.47	1.69	.99	-.15	.89
Nigeria					
1	-.94	.18	.96	-.51	1.14
2	.10	2.76*	.91	-1.66	1.45
3	1.96	.33	.99	-.08	1.04
4	.28	2.68*	.95	-.86	1.25
5	.32	.78	1.01	.14	1.02
6	.34	2.37	.94	-1.04	1.36
7	.25	1.01	.97	-.47	1.11
8	2.16	-.41	1.01	.18	1.00

*p < .01

[a]Weighted

say, each of these items, although it measured the same construct as the other items in the test, was inconsistent in measuring this single trait across different levels of intelligence within the same cultural group. Since this inconsistency might lead to the assignment of statistically unequivalent item parameter estimates to two samples that might be markedly different in ability on the construct measured, and this might be erroneously interpreted as bias on the part of the item, it was decided that all items with a significant between-group fit statistic be dropped from further analysis. This procedure might be considered an initial screening for potential bias.

Detection of Bias

Once the data sets had been cleared of irregular data, item difficulty estimates for each of the two cultural groups were compared using the rationale proposed by Wright (1977). Since estimates of item parameters now would not vary significantly unless the item differed in "cultural" loadings (Whitely & Dawis, 1974), any significant differences in item parameter estimates would indicate cultural bias in that item. The parameters for 22 items were judged to be similar for the two cultural groups and different (that is, "biased") for 11 items. In a parallel analysis, when similar comparisons were performed with the parameter estimates for two randomly formed groups, no sign of any bias was observed for any item, and this suggested that the cultural division was valid.

Six of the 11 items were seen to be "biased" in favor of the American group, while five were seen to be biased in favor of the Nigerian group. Those biased against the Nigerian group were Item 8 on Subtest 1, Items 2, 3, and 8 on Subtest 2, Item 10 in Subtest 3, and Item 3 on Subtest 4. Those biased against the American group were Item 10 on Subtest 1, Items 4, 7, and 8 on Subtest 3, and Item 1 on Subtest 4. The overall mean difference in difficulty parameter estimates was −.023; and although this was not significant, the difference was in favor of the American group.

For further analysis those 22 items identified to be unbiased were combined into a subset and termed "neutral" items, and those that favored the American group were combined into a subset of "American" items. Similarly, a subset of "Nigerian" items was formed.

All the CCFIT items were arranged in order of degree of fit, and the items in the first and fourth quartiles were isolated. With this division, a distribution of those items over the four subtests showed that Subtest 3 was the Rasch best-fitting subtest, whereas Subtest 2 was the Rasch poorest-fitting subtest. The KR20 reliability coefficient for the 22 neutral items for the American and the Nigerian groups, when corrected for test length, were .84 and .81, respectively. This showed a change, especially for the Nigerian group, from the original total test reliabilities of .82 and .75, respectively.

T tests done between the two groups on the set of neutral items, the set of American items, the set of Nigerian items, and the total CCFIT scores showed that the two groups differed significantly on total CCFIT scores and on both the sets of American and Nigerian items but did not differ in the sets of neutral items (see Table 6). With the total CCFIT scores, the difference was not substantial ($\omega^2 = .011$) and might have been the effect of large sample sizes. The American group scored significantly higher on both the total CCFIT scores and on the set of American items, while the Nigerian group scored significantly higher on the set of Nigerian items. These results indicate that the Rasch model could be used to select a set of (22) items that correlated less with culture than does the total test (see Table 7).

To give an idea of the expected range of the item difficulties within the present data, both the American and Nigerian samples were randomly divided in half, and item parameters were calculated for all four resulting samples; within each cultural group, differences in item difficulties were calculated for each item. For each item of the test, this within-culture difference was not significant. Table 8 presents the means and standard deviations for these data. The homogeneity of the differences and standard errors suggests that the results isolating biased items (Tables 2 through 5) were legitimate.

Discussion and Conclusions

Although a significant difference between the American and Nigerian groups was observed for the Rasch poorest-fitting subtest (Subtest 2) in favor of the American group, no such difference was found for the Rasch best-fitting subtest (Subtest 3). Whether or not a mean difference of .52 points is substantial enough to be concerned about is debatable. For Subtest 3 this means one-half point out of 14 questions. For the total test the significant difference was 1.03 points out of a total of 46. This is some indication that tests of significance, especially overall tests (Stelzl, 1979), may not be sensitive to differences. The data presented here, however internally consistent, might still have items with different characteristics that were undetected. The question that might be raised involves the degree to which a test can ever be perfectly unbiased.

The data do raise several further interesting questions. Cattell and Horn (1978) referred to the test as a perceptual test of intelligence as opposed to a verbal test. Subtest 2, that favoring the American children, is basically classifications ("Which of five stimuli does not belong?");

Table 6
T-test of Differences Between the Two Cultures in Total CCFIT Scores, "Neutral" Items, Rasch Best-Fitting Subtest, Rasch Worst-Fitting Subtest, "American" Items, and "Nigerian" Items

Subtest	Number of Items	Means	SD	t	p
Total CCFIT Items	46				
U.S.A.		31.18	6.18		
Nigeria		30.15	5.68	-3.95	<.001
"Neutral" Items	22				
U.S.A.		14.79	3.70		
Nigeria		14.49	3.19	-1.50	<.120
Rasch Best-Fitting Subtest 3	12				
U.S.A.		9.01	2.19		
Nigeria		8.96	2.14	-.45	<.625
Rasch Worst-Fitting Subtest 2	14				
U.S.A.		8.05	2.01		
Nigeria		7.53	1.94	-3.74	<.001
"American" Items	6				
U.S.A.		4.30	1.14		
Nigeria		3.58	1.10	-11.85	<.001
"Nigerian" Items	5				
U.S.A		3.09	1.16		
Nigeria		3.39	1.04	4.81	<.001
				df = 1401	

Table 7
Correlations Between Culture, "Neutral," "American," and "Nigerian" Items, Subtest 2, Subtest 3, and the Total CCFIT Scores

Variable	Variable 2	3	4	5	6	7
1. Culture[a]	-.06	-.29	.12	-.12	-.02	-.11
2. "Neutral" Items	--	.51	.60	.52	.76	.78
3. "American" Items		--	.36	.52	.44	.68
4. "Nigerian" Items			--	.27	.63	.62
5. Subtest 2				--	.36	.61
6. Subtest 3					--	.68
7. Total test scores						--

[a]Culture (U.S.A. = 1, Nigeria = 2)

Table 8
Means and Standard Deviations for Differences and the Resulting Z Ratios Between Item Difficulties for Randomly Created Subgroups Presented by Cultural Group and Subtest

	Cultural Group							
	America				Nigeria			
	Differences		Z Ratios		Differences		Z Ratios	
Subtest	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	-.04	.23	-.26	.98	-.04	.15	-.16	.71
2	-.15	.15	-.61	.78	.09	.26	.33	1.00
3	.12	.30	.17	.94	-.09	.19	-.16	.79
4	.11	.18	.58	.87	.07	.10	.33	.58

while Subtest 3, that favoring the Nigerian children, was progressive matrices. Of substantive interest is the question of the geometric nature of the stimuli and the classification nature of Subtest 2 compared with the deductive, extrapolative nature of Subtest 3. It might be valuable to determine if these differences replicate and then, if they do, to determine the extent they would replicate with other perceptual stimuli. If these differences are attributable to the differences between the English-style and American school systems, the effect of culture on intellegence would be far more subtle and indirect than the obvious but superficial influences of language, school content, and politics.

Piswanger (cited in Fischer & Pendl, 1980) described a similar conclusion: His sample of African students appeared to be identical to a group of Austrian students in progressive matrices tests except that the Africans performed more poorly when the changes involved horizontal movement. The author suggested that this was a result of the stress put on reading in Austria.

Compared to the other items of CCFIT, the set of those items that showed a better fit to the Rasch model—that is, provided a better measure of fluid intelligence—did not differentiate between the two groups. This observation supported Cattell's contention that uncontaminated fluid intelligence is not culture dependent. But the apparent overall lack of culture bias often observed for CCFIT might result from interitem neutralization of significant biases, in opposing directions, of most items in the test. The initial factor analysis, which (according to Jensen, 1980) would indicate measurement of a single criterion, did not indicate this. Jensen (1980) did state, however, that such biases that balance themselves out "cannot be claimed to measure one and the same ability on the same scale in both groups" (p. 445); this statement would then suggest the Rasch item selection procedure, compared to a factor analysis, would yield a test with less culturally loaded items.

References

Anderson, E. B. The numerical solution of a set of conditional estimation equations. *The Journal of the Royal Statistical Society, Series B,* 1972, *34,* 42-54.

Anderson, E. G. *Discrete statistical models with social science applications.* Amsterdam: North-Holland, 1980.

Biesheuvel, S. Adaptability: Its measurement and determinants. In L. J. Cronbach & P. J. D. Drenth (Eds.), *Mental tests and culture adaptation.* The Hague: Mouton, 1952.

Cattell, R. B. A culture free intelligence test I. *Journal of Educational Psychology,* 1940, *31,* 161-180.

Cattell, R. B. Are IQ tests intelligent? *Psychology Today,* 1968, *2,* 56-62.

Cattell, R. B. *Abilities: Their structure, growth, and action.* Boston: Houghton Mifflin, 1971.

Cattell, R. B. Are culture fair intelligence tests possible and necessary? *Journal of Research and Development in Education,* 1979, *12,* 2-13.

Cattell, R. B., & Horn, J. L. A check on the theory of fluid and crystallized intelligence with description of new subtest designs. *Journal of Educational Measurement,* 1978, *15,* 139-164.

Dague, P. Development, application and interpretation of tests for use in French-speaking black Africa and Madagascar. In L. J. Cronbach & P. J. D. Drenth (Eds.), *Mental tests and cultural adaptation.* The Hague: Mouton, 1972.

Davis, A. *Social class influence upon learning.* Cambridge MA.: Harvard University Press, 1948.

Davis, W. A., & Havighurst, R. J. The measurement of mental systems. *Science Monograph,* 1948, *66,* 301-316.

Drenth, P. J. D. The use of intelligence tests in developing countries. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology.* Amsterdam: Swet & Zeitlinger B. V., 1977.

Drenth, P. J. D., & Van der Flier, H. Cultural differences and comparability of test scores. *International Review of Applied Psychology,* 1976, *25,* 137-144.

Eckensberger, L. H. The necessity of a theory for applied cross-cultural research. In L. J. Cronbach & P. J. D. Drenth (Eds.), *Mental tests and cultural adaptation.* The Hague: Mouton, 1972.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. *Intelligence and cultural differences.* Chicago: University of Chicago Press, 1951.

Fischer, G. H., & Pendl, P. Individualized testing on the basis of the dichotomous Rasch model. In L. Van der Kamp, W. Langerak, & D. de Gruijter (Eds.), *Psychometrics for educational debates.* New York: Wiley, 1980.

Horn, J. L., & Cattell, R. B. Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology,* 1966, *57,* 253-270.

Hunt, M. McV. Heredity, environment, and class or ethnic differences. In *Assessment in a pluralistic society. Proceedings of the 1972 invitational conference on test problems.* Princeton NJ: Educational Testing Service, 1972.

Institute of Personality and Ability Testing (IPAT). *Measuring intelligence with the culture fair tests. Manual for Scales 2 and 3.* Chicago: Author, 1973.

Ironson, G. H. *A comparative analysis of several methods of assessing item bias.* Paper presented at the American Educational Research Association annual convention, Toronto, March 1978.

Iwuji, V. B. C. *An exploratory assessment of the effects of specific cultural and environmental factors on the performance of Nigerian students on selected verbal and nonverbal aptitude tests developed for use in Western countries.* Unpublished doctoral dissertation, University of Chicago, 1978.

Jensen, A. J. *Bias in mental testing.* New York: The Free Press, 1980.

Jordan, A. M. Parental occupation and children's intelligence scores. *Journal of Applied Psychology,* 1933, *17,* 103-119.

Leahy, A. M. Nature-nurture and intelligence. *Genetic Psychological Monograph,* 1935, *17,* 236-308.

Lord, F. M., & Novick, M. *Statistical theories of mental test scores.* Reading MA: Addison-Wesley, 1968.

Manaster, G. J., & Havighurst, R. J. Cross national research: Social-psychological methods and problems. Boston: Houghton Mifflin, 1972.

Massad, C. E. International scene. *Measurement News,* 1978, *9,* 3.

Mellenbergh, G. J. Applicability of the Rasch model in two cultures. In L. J. Cronbach, & P. J. D. Drenth (Eds.), *Mental tests and cultural adaptation.* The Hague: Mouton, 1972.

Neff, W. S. Socioeconomic status and intelligence: A critical survey. *Psychological Bulletin,* 1938, *35,* 727-757.

Rasch, G. *Probablistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research, 1960.

Stelzl, I. Ist der Modelltest des Rasch-Modells geeignet, Homogenitatshypothesen zu prufen? Ein

Bericht über Simulationsstudien mit inhomoginen Deten. [Can tests of significance for the Rasch model be used to test assumptions of homogeneity? A report about simulation studies using nonhomogeneous data.] *Zeitschrift für experimentelle and angewandte Psychologie,* 1979, *26,* 652-672.

Wellman, B. Growth in intelligence under differing school environments. *Journal of Experimental Education,* 1934, *3,* 59-83.

Wellman, B. Mental growth from pre-school to college. *Journal of Experimental Education,* 1937, *6,* 127-138.

Whitely, S. E., & Dawis, R. V. The nature of objectivity with the Rasch model. *Journal of Educational Measurement,* 1974, *11,* 163-178.

Wright, B. D. Solving measurement problems with the Rasch model. *Journal of Educational Measurement,* 1977, *14,* 97-116.

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

Wright, B. D., & Douglas, G. A. Conditional versus unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement,* 1977, *37,* 61-66.

Wright, B. D., & Mead, R. J. BICAL: Calibrating rating scales with the Rasch model (Research Memorandum 23A). Chicago: University of Chicago, Department of Education, Statistical Laboratory, 1978.

Wright, B. D., & Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement,* 1969, *29,* 23-48.

Author's Address

Send requests for reprints and further information to Thomas E. Dinero, Associate Professor, Educational Foundations, 406 White Hall, Kent State University, Kent OH 44242; or H. Johnson Nenty, Nigerian National Youth Service Corps at Federal Government College, PMB 5126, Port Harcourt, Rivers State, Nigeria.