A COMPARISON OF THE EFFECTS OF CONVENTIONAL TESTING AND TWO-STAGE TESTING PROCEDURES ON ITEM BIAS AS DEFINED BY THREE STATISTICAL TECHNIQUES

by

Carolyn Elizabeth Jones Lane

Dissertation submitted to the Graduate Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in

Educational Research and Evaluation

APPROVED: Lawrence H. Cross
Lawrence H. Cross, Chairman

Robert B. Frary
Robert B. Frary

Robert S. Schulman
Robert S. Schulman

Lawrence McCluskey
Lawrence A. McCluskey

L Wolfle
Lee M. Wolfle

October, 1978
Blacksburg, Virginia

To Dr. Dave, who kept his faith in the paradox.

ACKNOWLEDGMENTS

I would like to express my gratitude to my committee chairman, Dr. Lawrence H. Cross, for his many helpful suggestions and his effective assistance throughout the course of this research study. His willingness to give generously of his time and humor has been sincerely appreciated.

I also want to express appreciation to the members of my dissertation committee, Dr. Robert Frary, Dr. Lawrence McCluskey, Dr. Robert Schulman, and Dr. Lee Wolfle, for their assistance and constructive criticism throughout this study. I owe a special thanks to Dr. Robert Frary for introducing me to latent trait theory, and to all committee members for their willingness and interest in learning the principles of the Rasch model.

TABLE OF CONTENTS

CHAPTER Page

APPENDIX

LIST OF TABLES

TABLE Page

LIST OF FIGURES

FIGURES Page

Chapter I

STATEMENT OF THE PROBLEM

The elminiation of bias in testing is an essential educational measurement goal to ensure equal opportunity for all social groups. Measurement bias has become a topic for national debate due to recent litigation involving the use of tests for the selection of minority group members for employment and educational opportunities. Many maintain that traditional tests of academic achievement are biased in favor of a white middle-class culture and hence are inherently inappropriate for use with minority groups.

Despite the fact that bias has emerged as one of the most important and controversial topics in educational measurement, it has proven difficult to even define bias. Operational definitions have been divided into two general categories, external or internal, depending upon whether the investigations are based upon criteria external to the test scores (e.g., those investigating the extent to which a test predicts success in college) or whether the investigations are based solely upon information available from responses to the items. Furthermore, the levels of investigation have also differed; that is, the study of bias has been approached at both the item level and the test level.

Many deficiencies have been reported in traditional procedures for investigating test or item bias. It has been shown that different models of bias based on external criteria are contradictory (e.g., see Cole, 1973). Also complications such as bias in the criterion or different within-group reliabilities further complicate the task of the investigation of bias derived from the comparison of tests to an external criterion (Linn and Werts, 1971). Furthermore, criterion related validity data are usually not available during test construction, when it is important to identify and modify potentially biased items. Hence, for example, in a pilot study when external criteria are not available, it is important to study and develop statistical procedures for analyzing items based only on the information contained in the responses of persons to the items.

Approaches have included the comparison of traditional item difficulties (proportion of sample answering the item correctly), wherein a test is considered biased if the item difficulties differ for the groups of interest. The problem with this approach is that it is based on an unrealistic implicit assumption of identical ability distributions in the groups. An extension of this mean difference definition considers a test biased if there is significant group by item interaction as determined by a factorial analysis of variance (e.g., Cleary and Hilton, 1968). However, the significance tests resulting from these ANOVA designs may not be valid since the observations are non-normal, discrete random variables and the cell variances tend to be heterogeneous (Echternacht, 1974). While the underlying ability distribution may be considered normal, and hence continuous, the real problem remains the fact that the distribution of observed scores within the interval [0,1] may be skewed due to ceiling effects.

Approaches based on classical test theory have not in general been successful. The procedures fail to distinguish between differential test performance due to actual differences in ability between groups and differences due to measurement bias. Ability is an inherent characteristic of an examinee and should not be affected by test characteristics. Similarly, it should be possible to determine the difficulty of test items independent of the characteristics of the particular persons to whom the test is administered. However, methods for detecting bias based on the traditional definition of item difficulty fail to separate test characteristics from the ability distribution of the respondent sample. The traditional definition of item difficulty is sample dependent, and consequently bias definitions based thereon are also sample dependent. Hence, conclusions are relevent only to samples drawn at random from the same population used in the investigation and may not be applicable for groups with a different distribution of ability.

Alternative approaches to defining test bias at the item level are provided by item characteristic curve (icc) theory. The attractiveness of item characteristic curve theory, or latent trait theory, is that it permits the calibration of item parameters independent of the ability distribution of the sample (Hambleton et al., 1977). Furthermore, estimates of ability can be made independent of the set of items used for measurement. Hence, latent trait theory permits the separation of test characteristics from the ability distribution of the sample, an essential ingredient for the development of an objective definition of test bias (Hambleton et al., 1977).

A latent trait model specifies the relationship between observable examinee test performance and the unobservable trait or ability being measured by the test. This relationship between the observable and unobservable quantities is expressed by a mathematical function, the item characteristic curve (icc), which relates the probability of success on an item to the ability measured by the test.

The item characteristic curve represents the nonlinear regression of item score on the latent trait. The shape of the item characteristic curve does not depend upon the distribution of ability in the examinee population (Hambleton et al., 1977). As indicated in Figure 1, the same icc would be obtained regardless of whether group A or group B is used in the calibration of the curve, although obviously the ability distributions differ for the two groups. This invariance property of the item characteristic curves is one of the attractions of latent trait theory.

Latent trait theory also facilitates objectivity in measurement. Not only can item parameters be calibrated independently of the examinee sample, but also measurement of examinee ability can be made independently of the items administered. For example, two different, even nonparallel, tests consisting of items from a calibrated item pool may be administered to two students, and yet the resulting ability estimates for the two students are expressed in the same scale and are directly comparable. A test consisting of "easy" items may be administered to a student of low ability, a test of "hard" items to a student of high ability. The total score of each student on his respective test may be identical, but the corresponding latent ability estimates would differ. The mathematical properties of latent trait theory permit the separation of item difficulty from person ability so that comparisons of items can be distinguished from differences in examinee ability and vice versa.

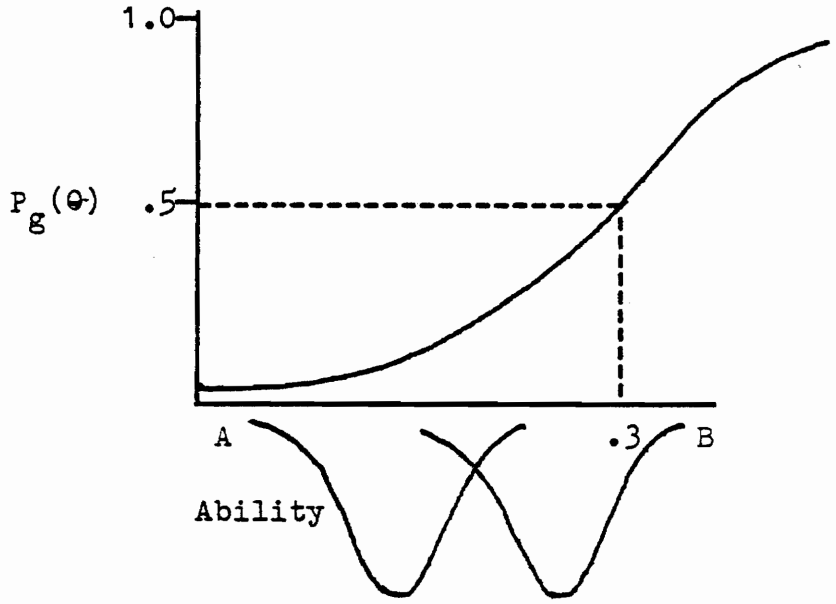


Figure 1

Sample Invariance of
Item Characteristic Curves

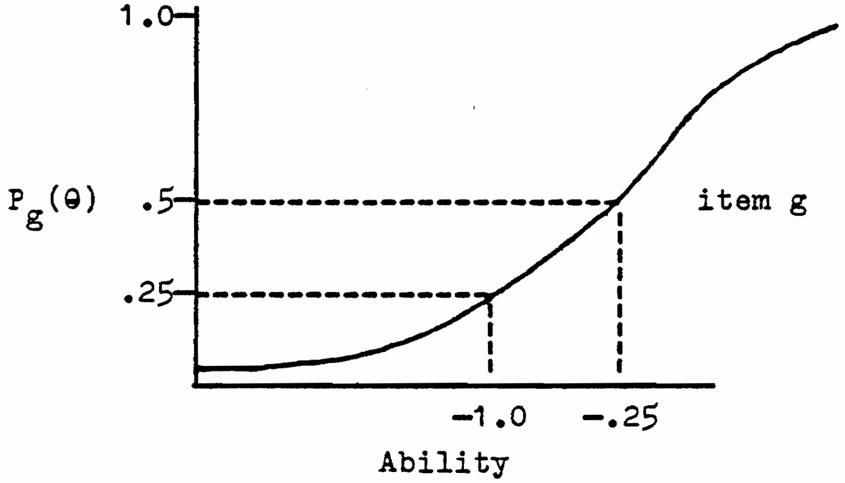


Figure 2

Example of an Item
Characteristic Curve

Several mathematical forms for the item characteristic curves have been employed, including the Rasch model. This one parameter model is based on the assumption that the probability of a correct response depends only on the difficulty of the item and the ability of the examinee. The Rasch model is the only model consistent with number right scoring (Wright, 1977b), and hence is the only model consistent with the scoring procedure most frequently applied with multiple-choice achievement tests.

According to this model the item characteristic curve has the form

$$P_g(\theta) = \frac{\exp(\theta - b_g)}{1+\exp(\theta - b_g)}$$

where θ represents the ability of the examinee and b_g the difficulty of item g. Ability, θ, is measured on the continuous scale $(-\infty, \infty)$, but usually θ is in the interval (-3, 3).

Figure 2 shows a graphical representation of an icc. Observe that for item g a person with ability -1.0 has a probability of .25 of responding correctly to this item. A person with ability -.25 has a probability of .5 of responding correctly to this item. Item difficulty is expressed on the same scale as ability. The difficulty of item g (b_g) is that ability level at which the probability of success is .5. Item g is thus said to have a difficulty of -.25, It should be noted that zero ability (θ=0) does not represent the absence of ability, but it often represents an average ability for a group.

During calibration an independent test of fit to the Rasch model is available for each item (Wright, 1977b). Durovic (1975) developed a definition of test bias based on a comparison of the tests of fit between groups. This definition results from the Rasch model's assumption that a test is unidimensional, i.e., that all items are measuring the same trait. The mean square fit for a given item indicates the extent to which that item is measuring the same trait as all the test items. Durovic's definition of bias indicates that an item is biased if it relates differently to the trait being measured for the two groups. By this definition an item that fits the model in a similar way for each group or one that fails to fit in a similar way for each group would not be considered biased. However, an item that fits the model for one group, but fails to fit for the other group, would be considered biased (Durovic, 1978).

Failure to fit the model would indicate that the item may involve another dimension. Perhaps a cultural factor is involved to which one group has been exposed, but which is outside the experience realm of the other group. In such a case, the item would be measuring not the trait desired, but the acquisition of this extraneous cultural factor.

Wright, Mead, and Draba (1976) and Mead (1976) utilized the Rasch model for the detection of bias through an analysis of residuals from the model. Mead (1976) proposed a graphical analysis of residuals. He investigated patterns of residuals plotted against the ability scale, which indicated various disturbances such as guessing, carelessness, speed and bias.

Draba (1977) further proposed a procedure for identifying biased items based on the assumption that the Rasch model produces statistically equivalent difficulty estimates for groups regardless of the ability distributions of the groups. The Rasch model is a probabilistic model; parameter estimates may fluctuate between groups, but the estimates should be statistically equivalent. A significant shift in the difficulty estimates of an item for the groups would indicate that the item was interacting with some characteristic of the group besides ability (Draba, 1977).

Each of these statistical definitions of bias only describes a technique for the detection of bias. Further research is needed to determine more definitive explanations of test bias and to establish guidelines for the construction of bias-free tests. The only solution to the problem of bias suggested by the procedures cited is the elimination of aberrant items from the item pool. Alternative strategies for the reduction or elimination of bias need to be carefully examined, including alternative testing procedures.

Recently, a number of adaptive testing strategies have been developed as alternatives to conventional testing procedures. In adaptive, or tailored testing, items are selected on an individual basis for each examinee. Tailored testing usually refers to a computer-interactive process, demanding the existence of a large pool of calibrated items and not lending itself easily to paper and pencil situations, making it impractical for most testing situations (Hambleton et al., 1977).

Another way of matching the difficulty of the items administered to the ability level of the examinee is to use a two-stage testing procedure. A two-stage testing procedure consists of a routing test followed by one of several second-stage tests, the choice of which is determined by the performance on the routing test. The second stage tests differ in difficulty, and students are directed to that test most appropriate to their ability estimate as determined by their score on the routing test.

Betz and Weiss (1974) compared the psychometric properties of two-stage adaptive tests and conventional tests. Their results indicate that two-stage testing better reproduces the true distribution of underlying ability than does conventional testing. Latent trait estimates obtained tended to be more reliable and more highly correlated with true ability. Their results indicated the potential superiority of two-stage testing over conventional procedures.

The potential superiority of adaptive testing in solving measurement problems such as bias needs to be investigated. Individualizing tests so that items are more appropriate to an examinee's ability level should reduce the effect of disturbances such as guessing on items that are too hard for an examinee or carelessness or boredom with items that are too easy. Furthermore, the possibility that the effect of an extraneous cultural variable present in an item may be minimal for examinees whose ability level is matched to the difficulty of the item should be investigated. It is hypothesized that individualizing tests for each examinee may have a positive effect in the direction of reducing or eliminating the number of items identified as biased. That is, it is proposed that often measurement error identified as bias is actually attributable to artifacts associated with inappropriate difficulty levels of tests. The goal of this study is to compare the effects of two-stage testing on item bias to the effects of the conventional testing procedure using a simulated two-stage testing procedure based on a real data set.

The primary analysis of item bias will be based on the Rasch latent trait model. Existing differences in ability should not affect the identification of item bias. Hence, the sample invariance property of the Rasch item parameters is considered most advantageous for identifying systematic error due to bias. Both Durovic's and Draba's definitions of bias will be considered.

Empirical support has been reported for both the Durovic and Draba procedures. Apparently, however, no empirical comparison of the procedures has been reported. Durovic's procedure, unlike Draba's, does not require that items fit the Rasch model. Unbiased items would fit or misfit the model in the same manner. Since Draba's definition evaluates the statistical equivalence of item difficulties across groups, it requires that items fit the Rasch model (Durovic, 1978). Durovic claims that it is likely that the two procedures may result in identifying different items as biased (Durovic, 1978). Comparisons will also be made with the traditional non-latent trait definition of bias originally proposed by Cleary and Hilton (1968).

In summary, this paper will compare the effects on item bias of conventional testing procedures to an adaptive testing procedure in which tests are individualized according to ability for each examinee by a two-stage testing procedure. The following two definitions of bias, both based on the Rasch latent trait model, will be considered:

1. An item is biased if it relates differently to the trait being measured for the two groups (Durovic, 1975).

2. An item is biased if its Rasch difficulty parameter differs significantly for the two groups (i.e., the item is interacting with some characteristic of the groups besides ability)(Draba, 1977).

The organization of this paper will be as follows: Chapter II will consist of a review of literature on bias studies based on classical and latent trait theory. Chapter III will contain a description of the data set and test to be used and the methods of analyses to be employed. It will include a discussion of the Rasch model, including estimation procedures to be used in the analysis, and a description of the two-stage testing procedure to be employed. Chapter IV will consist of a summary of the results of the analyses. Chapter V will conclude with a discussion of the results and conclusions as to any differential effects on bias evidenced in the comparison of the testing procedures.

Chapter II

REVIEW OF THE LITERATURE

Because of the increasing reliance upon testing, and because of possible discriminatory characteristics of some tests, there has recently been an increase in research on the nature and extent of test bias and test fairness. It has proven difficult to define what is meant by "biased" and "unfair". Bias refers to the psychometric properties of a set of test items or scores; test fairness is concerned with the way a test is used in a particular situation, such as in an employment or college selection procedure. Operational definitions of bias are divided into two general categories, depending upon whether the investigations are based upon external or internal criteria. The levels of investigation have also differed; the more common approach has been to consider entire tests rather than individual test items in the analyses.

Studies based on external criteria.

Methods employing external criteria generally involve the use of test scores for the prediction of some future criterion. The focus here is on the use of the scores rather than on the scores themselves.

One of the earliest definitions of test fairness in selection was stated in terms of significant differences between validity coefficients for the groups of interest (Pine, 1976). Any significant difference would imply that there was a discrepancy in prediction accuracy for the given subgroups. A lower test validity coefficient for a given subgroup is equivalent to a decrease in the variance of the predicted score distribution, which may lower the probability of selection for individuals in that subgroup (Pine, 1976).

Research continues on differential validity as a technique for determining test fairness; however, it is now recognized that equal validities are a necessary but not sufficient condition for test fairness (Pine, 1976). Consequently, other models have been proposed for defining selection bias. Five other major models of selection bias, as categorized by Cole (1973), will also be considered here: the quota model, the regression model, the subjective regression model, the constant ratio model, and the conditional probability model.

Social values may dictate that some group be favored in the selection process. In the quota model of bias the proportional representations of particular groups are established a priori on the basis of value judgments about fairness; failure to yield the specified proportions during the selection procedure is considered biased (Cole, 1973).

Definitions of test bias in the regression model emphasize consistent errors of prediction. Cleary (1968) provided a widely accepted definition of test bias which compares regression equations of criterion on test scores for different groups. She stated:

> A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias there may be a connotation of "unfair", particularly if the use of the test produces a prediction that is too low (Cleary, 1968, p.115).

A number of empirical studies (e.g., Cleary, 1968; Bower, 1970) have investigated possible bias in applying regression equations based on a majority group in the selection of minority group members for college admission. If the within group regression equations are not identical, then separate within group regression equations should be employed according to this definition of fairness (Cleary, 1968). Therefore the application of Cleary's definition is operationally equivalent to endorsing differential prediction in selection (Pine, 1976).

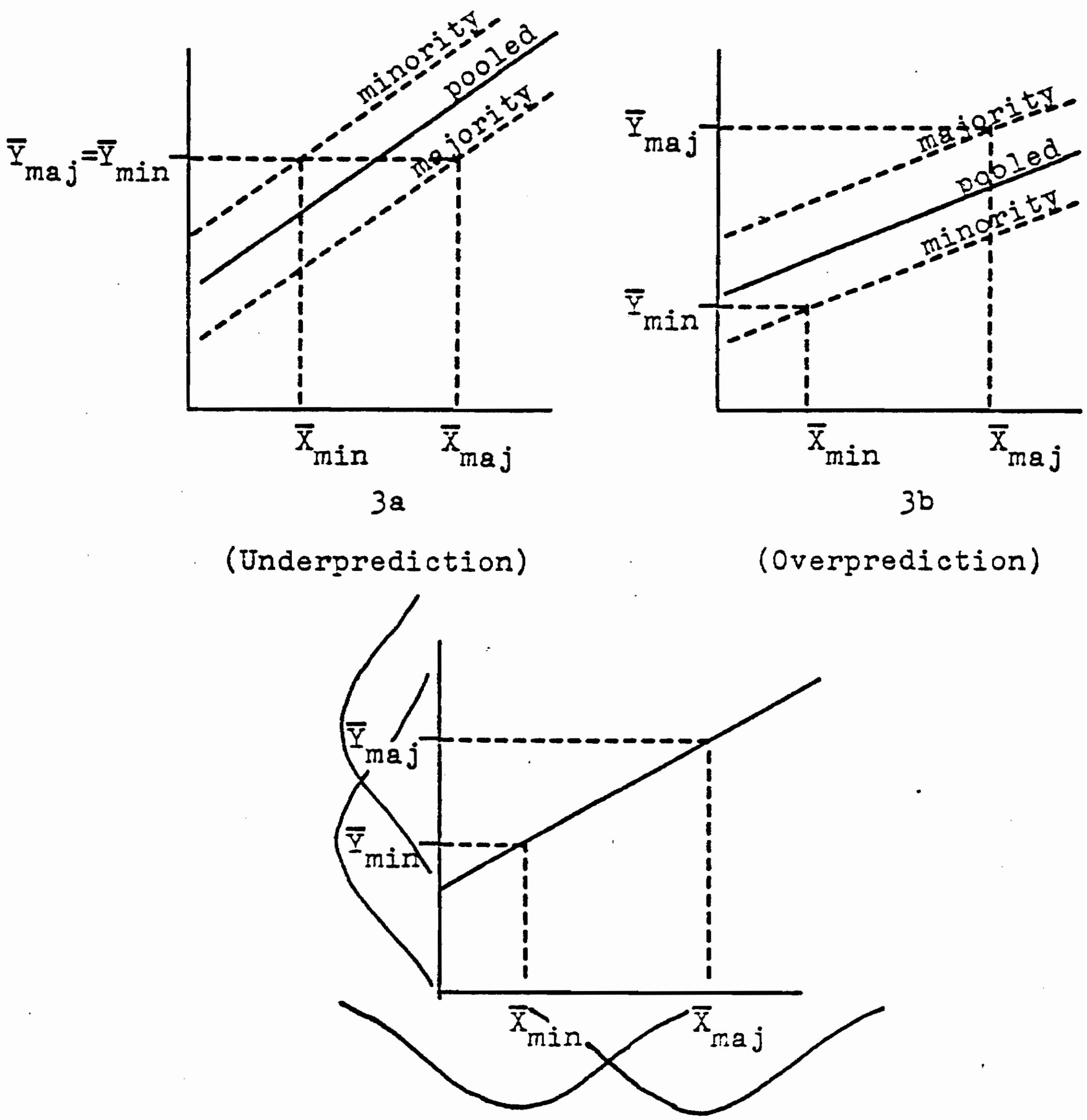
Consider the situation depicted in Figure 3, in which it is assumed that the within group regression slopes for the majority and minority groups are equal. In Figure 3a the mean criterion score for the minority subgroup is equal to the mean criterion score for the majority subgroup ($\bar{Y}_{maj}=\bar{Y}_{min}$), but the mean predictor test score of the majority subgroup is greater than the mean score of the minority subgroup ($\bar{X}_{maj}>\bar{X}_{min}$). Here either the regression line of the majority subgroup or the pooled regression line would result in underprediction of the minority subgroup. In Figure 3b the groups differ on both predictor and criterion mean scores by 1 SD. In this case using either the majority or the pooled regression line would result in overprediction for the minority subgroup.

Darlington (1971) translated definitions of cultural fairness into correlational terminology, using a cultural variable C, usually considered as dichotomous. He restated Cleary's definition as

> Test X is culturally fair if knowledge of a person's cultural group cannot be used to improve the prediction of Y made from X. (Darlington, 1971, p.73).

In correlation terms

$$\rho_{CY \cdot X} = 0.$$



3a

(Underprediction)

3b

(Overprediction)

3c

(Identical Regression Lines)

Figure 3

Effect of Different Regression Lines
on Prediction for Minority Groups

Thorndike (1971) demonstrated that a test might be fair by Cleary's definition, but unfair by another standard. He considered examples as in Figure 3 in which the slopes of the within group regression lines were equal, but in which the minority group differed from the majority group in its mean score on the predictor test (X), the criterion measure (Y), or both.

Consider Figure 3c in which the predictor mean score for the majority subgroup is 1 SD greater than the score for the minority subgroup, and the majority criterion score is 1/2 SD greater than the minority criterion score, and in which the regression lines actually coincide. By Cleary's definition this would be a fair test. Assume that the distributions are normal and that the variances are equal for the two subgroups, and suppose that the majority mean score is the selection cut-off. Then only about 16 percent of the minority group would qualify for selection, whereas about 31 percent would actually exceed the critical value on the criterion measure. Hence Thorndike (1971) proposed an alternative definition of bias, the constant ratio criterion:

> ...the qualifying scores on a test should be set at levels that will qualify applicants in the two groups in proportion to the fraction of the two groups reaching a specified level of criterion performance (Thorndike, 1971, p.63).

Thus Thorndike's definition states that the success ratio should equal the selection ratio, that is

$$\frac{Pr_{min}\ (Y>Y_p)}{Pr_{maj}\ (Y>Y_p)} = \frac{Pr_{min}\ (X>X_{min})}{Pr_{maj}\ (X>X\ maj)}$$

where Y_p is the determined critical level for the criterion, and X_i

the selection critical levels on the predictor variable for the two subgroups. In correlational terms, Darlington (1971) stated that Thorndike's definition implies that

$$\rho_{CX} = \rho_{CY}$$

Darlington (1971) argued that a subjective decision must be made concerning the relative importance of the two goals of maximizing validity and minimizing discrimination. He suggested that the data be used to specify a linear function, Y-kC, of the criterion Y and a cultural variable C, usually dichotomous, to be maximized for "cultural optimality". He urged that.......

> the term "cultural fairness" be replaced in public discussions by the concept of "cultural optimality". The question of whether a test is culturally optimum can be divided in two: a subjective, policy-level question concerning the optimum balance between criterion performance and cultural factors (operationalized in our equations as the optimum value of k), and a purely empirical question concerning the test's correlation with the culture-modified criterion variable (Y-kC) and whether that correlation can be raised. Anyone who objects to a test on the latter ground can be invited to construct a test correlating higher with (Y-kC) (Darlington, 1971, p. 79).

Hence, like the quota model, the subjective regression model requires a value judgment on the importance of the selection of members of some group over another.

Cole (1973) noted that in many models of selection bias it may happen that the probability of selection of potentially successful minority group members is different from the probability of selection of such majority group members. That is, the probability of selection depended upon group membership. Cole proposed a model which eliminates this type of unfairness:

> The basic principle of the conditional probability selection model is that for both majority and minority groups whose members can achieve a satisfactory criterion score ($Y > Y_p$) there should be the same probability of selection regardless of group membership. (Cole, 1973, p. 240).

In probability terms

$$Pr_{maj}\{X>X_{maj} \mid Y>Y_p\} = Pr_{min} \{X>X_{min} \mid Y>Y_p\}$$

In correlational terms, the conditional probability model satisfies Darlington's definition (1971) requiring that

$$\rho_{CX \cdot Y} = 0.$$

Studies based on internal criteria.

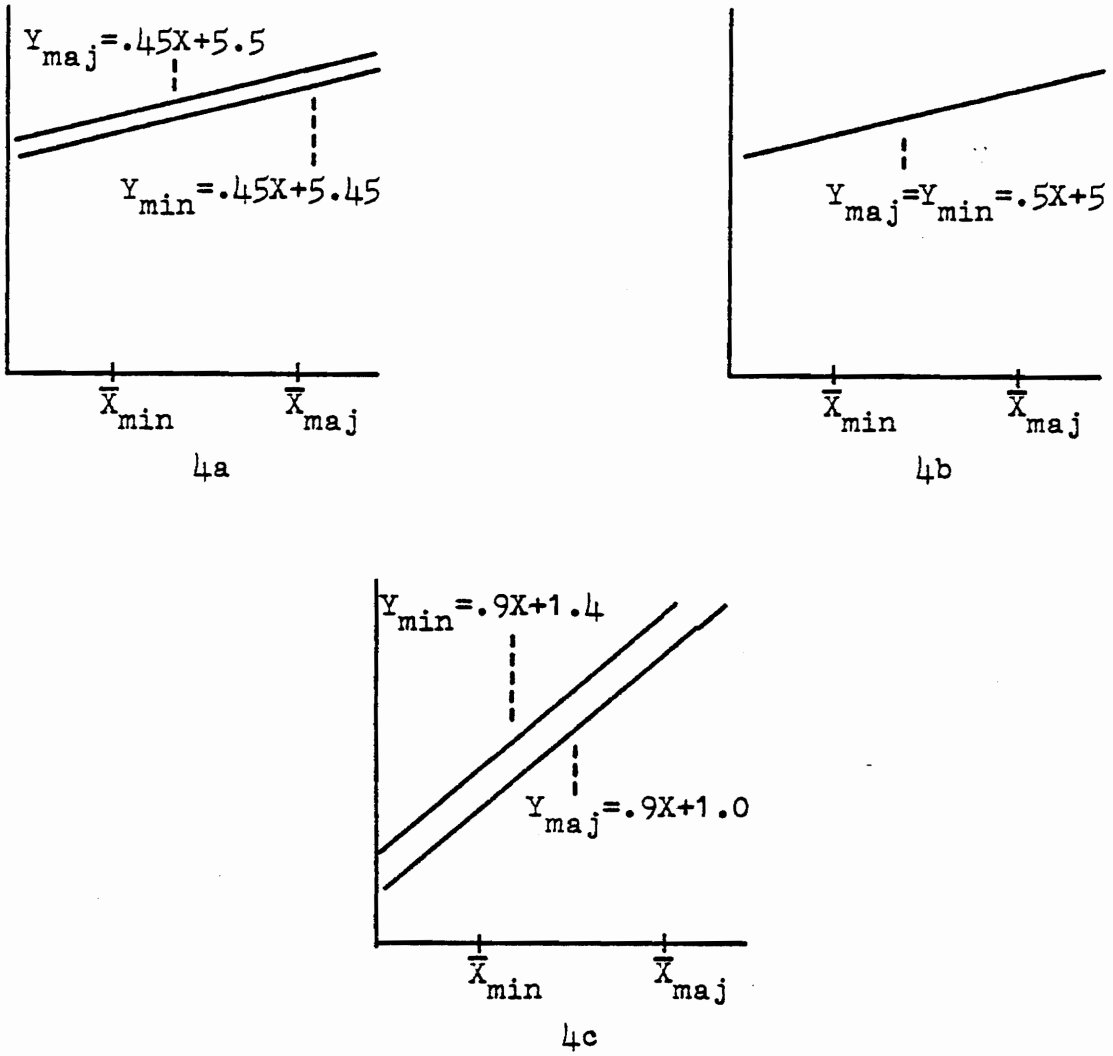
Darlington (1971), Thorndike (1971), and Cole (1973) showed by considering realistic hypothetical cases that the above models of bias in selection are contradictory. Linn and Werts (1971) also noted that complications such as bias in the criterion and unrealiability of the predictors further complicate the task of the investigation of bias derived from the comparison of tests to an outside criterion.

This is illustrated by an example described in Table 1 in which the validity coefficient and standard deviations for both the predictors and the criterion are equal for the majority and the minority group. The means, however, are not equal, resulting in regression lines with equal slope but slightly different intercepts. This situation is depicted in Figure 4a in which the "majority" equation overpredicts for the minority group. Since the tests are not perfectly reliable, given classical test theory assumptions, the true slope would equal the observed score slope divided by the reliability of the test. The "corrected" results, for a reliability of .9 for the data in Table 1, (as in Figure 4b), would result in underprediction for the minority

TABLE 1

EFFECT OF UNRELIABILITY OF THE PREDICTOR ON REGRESSION EQUATIONS

Group	Test	Mean	SD	Correlation of Predicts and Criterion	Within group Reliability Of Predictor	"Corrected" Slopes	Inter-cepts
Maj.	Predictor	10	1	.45	.9	.5	5.0
	Criterion	10	1		.5	.9	1.0
					1.00	.45	5.5
Min.	Predictor	9	1	.45	.9	.5	5.0
	Criterion	9.5	1		.5	.9	1.4
					1.00	.45	5.45



4a

4b

4c

Figure 4

Regression Lines for Different Predictor Reliabilities according to Thorndoke's (1971) definition. A reliability of less than .9 would result in underprediction for the minority group (Figure 4c); a reliability of greater than .9 would result in overprediction (Figure similar to 4a). Furthermore, criterion-related validity data are usually not available during test construction, when it is important to identify and modify potentially biased items. An alternative approach to the concept of test bias would involve models which made some statistical statements about the items in a test only from the information contained in the responses of persons to the items. One limitation to this approach is that it would fail to detect bias in the situation where all items in a test are biased.

The simplest definition of test bias in the absence of an external criterion states that a test is fair if there is no difference in mean total scores between populations A and B (Potthoff, 1966). This definition rules out a priori any differences in ability distributions between populations A and B. This assumption is not feasible given the current status of unequal schooling and cultural opportunities, and so differences in score distributions do not, per se, constitute evidence of bias. "It is not so much that the test is biased, but that there is bias in the learning environment that helps determine the test score" (Hambleton et al., 1977, p. 92).

At the item level, the mean difference definition states that a dichotomously scored item is unbiased if there is no difference in item difficulty, i.e., $p_{g,maj}=p_{g,min}$, where $p_{g,i}$ is the proportion of examinees in population i responding correctly to item g.

An alternative definition, expanding upon the mean difference definition, but not requiring an a priori assumption of identical ability distributions, states that a test is unbiased if there is no item by group interaction.

> An item on a test is said to be biased for members of a particular group if, on that item, the members of the group obtain an average score which differs from other groups by more or less than expected from performance on other items of the same test. That is, the biased item produces an uncommon discrepancy between the performance of members of the group and members of other groups. In terms of the analysis of variance, bias is defined as an item x group interaction (Cleary & Hilton, 1968, p.61).

Cardall and Coffman (1964) first applied the two-way ANOVA design to the study of bias, with the primary hypothesis of interest being the detection of item by group interaction. That is, they wished to determine if some items were relatively easier for one group than for others. Cardall and Coffman drew three samples (n=300) from each of three groups: Group I, rural midwestern whites; Group II, northern urban whites, Group III, southern blacks. Two ANOVAs were performed, one for the 40 verbal items from the May, 1963 SAT administration, the other for 25 math items. Significant group main effects were found. Interaction between groups and items was found to be significant, particularly on the math ANOVA.

Cleary and Hilton (1968) followed the pattern described by Cardall and Coffman (1964) to study the variation in item scores in different racial and socioeconomic (SES) groups. A three factor ANOVA was used. The first factor was race; the second SES, which was considered nested within race to avoid the assumption that SES levels are comparable across races; the third factor was items and was considered random. The dependent variable was item score on selected items of the 1961 and 1963 administrations of the PSAT. Again, the hypotheses of primary interest pertained to the existence of interaction. In particular, did item scores change as a function of race, SES within race, or perhaps both? Almost all tested effects were found to be significant, but this was perhaps attributable to the large within cell sample sizes (for 1961, n=106; for 1963 n=129). In view of such large sample sizes, the percentage contribution of each effect to the total variance was considered. Item by race (indicator of racial bias) and item by SES within race (indicator of SES bias) interactions each contributed minimally to total variance (less than 1%). Bivariate plots of sums of item scores indicated little deviation from linearity except possibly that caused by an apparent "floor" effect in black scores, which would contribute to a significant item by race interaction (Cleary and Hilton, 1968).

This "floor" effect is a drawback of any mean-difference procedure. Items at the extremes of the difficulty continuum may appear biased simply because of a floor or ceiling effect.

Continuing in the pattern originally established by Cardall and Coffman (1964), Angoff and Ford (1973) considered the effect on item by race interaction of matching the groups on some concurrent variable. Item p-values for each group were transformed to normal deviates and then to delta values by the linear transformation $\Delta=4z + 13$. The delta values for blacks were plotted against the delta values for whites. The correlation coefficient represented by the plot of the deltas for the two groups was used as an expression of item by race interaction. Their findings indicated that matching the groups on a related ability (the groups were matched on verbal PSAT scores to study math scores and vice versa) decreased part of the disparity observed in the item by race interaction.

An objective of an ETS study reported by Breland et al. (1974) was to explore the problem of bias with a combined statistical and subjective approach. It is their view that "...given the importance of the use to which items (or entire tests) are put, no entirely mechanical procedure would seem likely to gain acceptance" (Breland et al., 1974, p. 4). Potentially biased items were detected using the procedure of Angoff and Ford (1973), and the subjective analyses were applied to determine any peculiar characteristics of these aberrant items (Breland et al., 1974).

ANOVA may not be a valid design for the study of bias. Ecternacht (1974, p. 272) noted that "the significance tests resulting from this type of analysis are somewhat suspect, in that the observations are non-normal discrete random variables and the cell variances tend to be non-homogeneous." He consequently developed an alternative procedure for the detection of bias, based on the item by group interaction definition, which corrects for a "ceiling" effect of items at the extremes of the difficulty continuum. He defines a test to be absent of bias (item by group interaction) if the differences $p_{ij}-p_{ij}'$, where p_{ij} denotes the difficulty of item i,i=1,2,...n, for group j,j=1,2...m, are constant over all n items for groups j,j'. His technique transforms sample item difficulties p_{ij}, to delta values, Δ_{ij}, using the inverse normal transformation and the linear transformation $\Delta_{ij} = 13\text{-}4z_{p_{ij}}$, where $z_{p_{ij}}$ denotes the value corresponding to a cumulative normal ordinate value of p_{ij}. Echternacht states that

> Under a hypothesis of no item-group interaction, the sample differences $\Delta_{ij}\text{-}\Delta_{ij'}$ should be distributed normally with some unknown mean and variance for each group j and j'. If evidence can be gathered to the effect that this is not the case, the null hypothesis can be rejected (Echternacht, 1974, p. 274).

Basically, he used a modification of the Kolmogorov-Smirnov test for normality where the hypothesized mean and variance are calculated from the sample data. His procedure is contingent upon the acceptance of the notion of item by group interaction as equivalent to test bias.

Scheuneman (1976) proposed a technique for the detection of item bias which requires no normality assumption and does not require representative samples. Her definition of bias states:

> An item is unbiased if, for all individuals belonging to the same ability group as defined by the total score on the test or subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered. Once the ability groups have been defined, a modified chi square procedure is used to evaluate each item in the test for possible bias (Scheuneman, 1976, p.1).

This method has been used to screen the item pool for the Metropolitan Reading Tests. Empirical results have been found to depend upon the choice of ability intervals in the contingency table.

Lord (1976) noted that matching on a test composed of items that are to be studied, such as suggested by Scheuneman, may introduce spurious relationships. Matching on parallel forms is likewise not advised, since matching would be done on fallible scores rather than on true scores, introducing a regression effect (Lord, 1976).

Studies based on latent trait theory.

Alternative approaches to defining test bias at the item level are provided by item characteristic curve theory. The attractiveness of item characteristic curve theory, or latent trait theory, is that it permits calibration of item parameters independent of the ability distribution of the sample.

> A theory of latent traits supposes that examinee performance on a test can be predicted (or explained) by defining examinee characteristics, referred to as traits, estimating scores on these traits and using the scores to predict or explain test performance (Lord and Novick, 1968, p. 358).

A latent trait model specifies the relationship between observable test performance and unobservable traits or abilities. The item characteristic curve expresses mathematically this relationship between the observable and unobservable quantities. The item characteristic curve (icc) relates the probability of success on an item to the underlying ability measured by the test that contains it; the icc represents the nonlinear regression function of item score on the latent trait measured by the test. "In other words, the shape of the item characteristic curve does not depend upon the distribution of ability in the examinee population. This invariance property of icc's and consequently the parameters describing the curves is one of the most attractive characteristics of latent trait models" (Hambleton et al., 1976, p.13). Several mathematical forms for icc's have been employed, e.g., normal ogive and logistic models. The appropriateness of the choice is difficult to validate because icc's represent the regression of item scores on a trait which is not directly measurable. Several investigations (e.g., Hambleton and Traub, 1971, 1973) have considered the "predictive" appropriateness of the logistic models with favorable results.

Birnbaum (1968) originally substituted the two-parameter logistic cumulative distribution function for the normal ogive item characteristic function because of simplistic mathematical properties of the logistic function.

The two-parameter logistic model has the form

$$P_g(\Theta)=\frac{\exp[Da_g(\Theta-b_g)]}{1+\exp[Da_g(\Theta-b_g)]}$$

where $P_g(\Theta)$ represents the probability that an individual of ability Θ will respond correctly to item g, which is scored dichotomously. D is a scaling factor; usually D=1.7 to obtain close agreement with the normal ogive model (see Birnbaum, 1968, p. 399). The parameter b_g is usually referred to as the index of item difficulty; it represents the point on the ability scale at which the examinee has a probability of .5 of answering the item correctly. The paramemter a_g is called the item discrimination and is proportional to the slope of $P_g(\Theta)$ at the point $\Theta=b_g$.

The two-parameter model implicitly assumes that guessing does not occur. Birnbaum (1968) proposed that a third parameter, c_g, commonly referred to as the guessing parameter, be included in the model to help account for the lack of fit of the icc at the lower end of the ability continuum, where guessing is more likely to occur. The mathematical form of the three-parameter model is

$$P_g(\Theta) = C_g + (1-C_g) \frac{\exp[Da_g(\Theta-b_g)]}{1+\exp\,[Da_g(\Theta-b_g)]}$$

Lord (1976) described a design developed by Marco at ETS for the detection of item bias using icc techniques. Due to the sample invariance property of latent trait theory, the icc for an item should be the same for all cultural groups. Lord advanced a two step procedure for the detection of biased items:

> 1. Plot latent item difficulties for one subgroup against item difficulty for the other subgroup to place all parameters on a common scale. The values of b_g for the subgroups should differ only in origin and unit of measurement. Hence the points should fit a straight line.
>
> 2. Test the hypothesis that items which demonstrated significant departures from linearity have the same icc for the subgroups of interest.

Pine (1976) used the normal ogive model to study the relationship between the item characteristics of a test and models of fairness. Pine defined bias as $b_{g\ maj} - b_{g\ min}$, that is, the difference in the latent trait difficulty parameters for the majority and the minority subgroups. In a Monte Carlo simulation he considered the effect of the distribution of item difficulty, i.e. peaked or uniform, the level of item discrimination, and the test length on Cleary's (1968) regression model of test fairness and on Thorndike's (1971) constant ratio model, using the true latent ability distribution as the "external" criterion. He found that item characteristics greatly affected the predicted ability distribution, with very complex relationships manifested between the variables.

The one-parameter logistic model was developed independently of the two-and three-parameter models by the Danish mathematician G. Rasch (1966). Essentially, it is a special case of Birnbaum's three-parameter model in which all items are assumed to have equal discriminating power, $\bar{a}$, which implies that the items differ only in difficulty. The equation for this model can be written

$$P_g(\Theta)=\frac{\exp[D\bar{a}\ (\Theta-b_g)]}{1+\exp[D\bar{a}(\Theta-b_g)]}$$

"The Rasch model follows from the assumption that the unweighted sum of right answers given by a person will contain all of the information needed to measure that person, and that the unweighted sum of right answers given to an item will contain all of the information needed to calibrate that item. The Rasch model, consequently, is the only model that is consistent with number right scoring" (Wright, 1977a, p. 102). In defense of the Rasch model over the two-parameter logistic model, Wright further states:

> if we want to think that the probability of success on the harder of two items should always be less than the probability of success on the easier, no matter who attempts the items, if that is what we intend by "harder", then we must see to it that variation in item discrimination sufficient to produce item characteristic curves that cross does not occur (Wright, 1977b, p. 103).

During calibration, an independent test of fit to the Rasch model is available for each item (Wright and Panchapakesan, 1969, p. 44). Durovic (1975) based his definition of test bias on this test of fit:

> an item is biased for members of a group, if on that item, for members of the group, a mean square

> fit of the item to the Rasch model is obtained which differs, by greater than one, from the mean square fit obtained for members of the other group. By this definition, a test is not biased if each item in the test relates to the dimension being measured in the same way for each group (Durovic, 1975, p. 4).

Durovic noted that since the Rasch model is a probabilistic model, perfect fit to the model is not expected. However, his criterion for determining biased items of differences in fit mean squares of one was an arbitrary decision based upon his research experiences (personal communication).

While empirically investigating the potential of his definition, Durovic explored the possibility of content-based explanations for misfitting items. Two reviewers with expertise in black cultural characteristics were selected to provide content evaluation. Their remarks supported the results obtained by the proposed definition (Durovic, 1975).

Wright, Mead and Draba (1976) and Mead (1976) utilized the Rasch model for the detection of bias through a systematic analysis of residuals from the model. The analysis of fit is an analysis of the discrepancy between the observed outcome and the outcome expected according to the model (See Appendix A for definition of residual). The proportion residual is transformed into an approximate logistic residual (Wright et al., 1976, p. 13). Wright et al. set up a weighted least squares ANOVA for testing differences in item difficulty between subgroups. One advantage of this approach cited is the opportunity to evaluate a variety of specific hypotheses about the nature of bias. Analogously to the test of fit, residuals of each person-item interaction may be used to analyze the appropriateness of items for each individual, making it theoretically possible to identify items that are fair for individuals as well as for specific subgroups. Mead (1976) proposed a graphical analysis of residuals. He investigated patterns of residuals, transformed to the ability metric, plotted against the ability scale. Various disturbances, such as guessing, carelessness, speed and bias, are indicated by the graphical patterns of residuals (Mead, 1976). The response pattern for each person can similarly be analyzed by evaluating the way in which these residuals correlate with difficulty, item position, and type (Wright, 1977b).

Draba (1977) further proposed a procedure for identifying biased items using the Rasch model:

> ...difficulty estimates should be statistically equivalent for groups distinguished only by their ability distributions. If, on the contrary, difficulty estimates shift significantly from group to group, this suggests that the item interacts with particular characteristics of the groups such as race or sex (Draba, 1977, p. 1).

Recall that the Rasch difficulty parameters should be group independent. The Rasch model is a probabilistic model; parameter estimates may fluctuate between groups but the estimates should be statistically equivalent. Draba used a t-test for testing the equivalence of the difficulty estimates. A significant shift would indicate that the item was interacting with some characteristic of the group besides ability (Draba, 1977). Recall that this is the definition of bias employed by Pine (1976), who based his study on the normal ogive model.

While empirical support has been reported for both procedures, apparently no empirical comparison of the procedures has been reported. Durovic claims that it is likely that the two procedures may result in identifying different items as biased (Durovic, 1978).

Adaptive testing procedures.

The major limitation of all these statistical definitions of bias is that each only describes a technique for the detection of bias, none suggest means for correcting the problem, except eliminating aberrant items from the item pool. Hypotheses need to be tested about why items are biased, and guidelines based on these results need to be established for the construction of unbiased items (Burrill, 1975). Furthermore, the effect of various testing procedures on bias needs to be analyzed. Recently, a number of adaptive testing models have been developed as alternatives to conventional testing procedures. In adaptive testing, items are selected on an individual basis for each examinee. If the group to be tested is relatively heterogeneous, it is impossible for a conventional test to measure accurately across the ability continuum. The difficulty levels of the items need to be matched to the ability level of the examinee to obtain effective measurement.

> When number right scores or conventional formula scores are used, the hard items not only waste the time of the low ability examinees, they impair whatever measurement of these examiness would otherwise be effected by the easy items....To obtain effective measurement at low (high) ability levels, we need easy (hard) items (Lord, 1977, p. 125).

Significant advances in the area of adaptive or tailored testing have been made through applications of latent trait theory. The invariant properties of latent trait item parameters permit the measurement of person ability independent of the particular items used for testing. Ability measures are on a common scale, facilitating the comparison of persons even when they have been measured with entirely different tests (e.g. see Wright, 1967; Forster, 1977; Lord, 1977).

Tailored testing usually refers to a computer-interactive process in which the computer has been programmed for the following repeated basic steps:

> 1. Before the next item is selected and administered, examinee ability from responses up to this point are estimated.
>
> 2. The item from those not yet administered is selected which is likely to measure most effectively at the examinee's estimated ability level.

The procedure is terminated when the desired measurement precision is attained (Lord, 1977). This procedure demands the existence of a large pool of calibrated items and does not lend itself easily to paper and pencil testing situations, making it impractical in most testing situations (Hambleton et al., 1977).

Another way of matching the difficulty of the items administered to the ability of the examinees is by using a two-stage testing procedure. A two-stage testing procedure consists of a routing test followed by one of several alternative second-stage tests, the choice of which is determined by the examinee's performance on the routing test (Lord, 1971).

Decisions must be made as to the appropriate length and difficulty level for each stage of the testing procedure. Lord (1971) noted that:

> the choice of difficulty level for each second stage test and the choice of the corresponding cutting scores on the routing test are most difficult matters. An unfortunate choice of cutting scores may lead to a situation where most examinees are routed to the same second-stage...or to other difficulties (Lord, 1971, p. 235).

If the routing test is too long, not enough items are left for the second stage tests; if it is too short, then examinees are poorly directed to the second stage tests (Lord, 1971). The routing test may be self-scored or even given in advance of the regular testing as a take-home self-scored test (Lord, 1977).

An empirical study of two-stage testing by Angoff and Huddleston (1958) indicated that the two-stage procedure was technically superior to a conventional test. Results showed that the two-stage tests tended to be more reliable in the subgroups for which they were intended than were conventional tests, and that predictive validities, with grade-point average as the criterion, were also slightly higher than those of the conventional test.

A series of studies reported by Linn, Rock and Cleary (1969) indicated that the majority of the two-stage procedures considered had higher predictive validities than did conventional tests of the same or longer length.

Betz and Weiss (1973, 1974) compared the psychometric characteristics of two-stage adaptive tests and conventional tests using both empirical and simulated techniques. Their results indicated that scores obtained from two-stage testing better reproduce the distribution of underlying ability. Latent estimates obtained were more reliable and were more highly correlated with the underlying ability than were the conventional test scores. Their results indicate that good design in the two-stage testing procedure can result in more information at all levels of ability than conventional tests. Betz and Weiss (1974) interpreted their results as indicating the potential superiority of two-stage testing over conventional testing procedures.

The potential superiority of the two-stage testing procedure over conventional testing in solving measurement problems needs to be further studied. The purpose here will be to consider the effect of two-stage testing on test bias. Angoff and Ford suggested that:

> ...it would appear that one element of the cultural difference between the two ethnic groups is the simple fact of the difference in levels of performance, and that the so-called cultural difference between the two races would diminish considerably if the performance levels were more similar (Angoff and Ford, 1973, p. 103).

Recall that Scheuneman (1976) attempted to circumvent the problem of disparity in underlying ability levels by defining item bias as unequal item difficulties within the same ability groups.

It is proposed that individualizing tests for each examinee may have some positive effects in the direction of reducing or eliminating the number of items identified as biased. Comparisons of the results of two-stage and conventional testing procedures should be considered using various definitions of bias.

The effects of disturbances such as guessing on items that are too difficult or carelessness on items that are too easy should be reduced by a testing procedure in which test items are appropriate to an examinee's ability level. Furthermore, such testing procedures may reduce the effect of extraneous cultural variables. Hence, it is proposed that individualizing tests for each examinee may have a positive effect in the direction of reducing or eliminating the number of items identified as biased. That is, it is believed that inappropriate test level explains much of the measurement error that is identified as bias. The goal of this study is to compare the effects of a conventional testing procedure with a simulated two-stage testing procedure based on a real data set.

Chapter III

METHOD

Rasch model

Group differences in performance on a test item can be due to actual differences in the underlying ability or to systematic error attributable to bias. The objectivity of the Rasch model permits the separation of test item characteristics from the ability distribution of the sample (Wright, 1969), and hence permits the avoidance of confusing performance differences due to differential ability levels from that due to cultural bias.

The Rasch model specifies one person parameter and one item parameter and defines the probability of a successful response to an item as an exponential function of these parameters.

> The Rasch model follows from the assumption that the unweighted sum of right answers given by a person will contain all of the information needed to calibrate that item. The Rasch model is the only model consistent with "number right" scoring (Wright, 1977b, p. 102).

Furthermore, the model is consistent with the notions that:

> 1. A more able person always has a better chance of success on an item than does a less able person.
>
> 2. A person has a better chance of success on an easy item than on a difficult one (Wright, Mead, and Draba, 1976).

Other latent trait models employ additional item parameters. A parameter for variation in the slope of the item characteristic curves may be introduced to allow differences in discrimination power of items. Consider for example Figure 5. Models which involve a discrimination parameter permit variation in slope and hence crossed item characteristic curves as shown. A person of ability Θ_1 would find item h easier than item g; however, an examinee of ability Θ_2 would consider item g easier. Consequently, a model consistent with the notions above must not allow for variation in item discrimination (Wright, 1977).

Another parameter is often included to allow variation in the lower asymptote of the item characteristic curve to account for guessing (see Hambleton et al., 1977).

The Rasch model is a special case of the two and three parameter models. However, it is easier to work with since it involves fewer parameters and since the problem of parameter estimation is essentially solved (Hambleton et al., 1977). Statistical and numerical problems exist in the estimation of parameters for the two and three parameter models. Computer programs are available for estimation, but "the method usually does not converge properly" (Lord, 1968, p. 1015). Furthermore, Hambleton and Traub (1973) in a comparison of the one- and two-parameter models noted:

> The results of the study suggest that the two-parameter test will provide greatest improvement over the one-parameter model when applied to data from short tests when the variability of the discrimination parameter is substantial. Whether the gains are worth the increased cost of solving for the parameters of the more complex method is a question which requires investigation (Hambleton and Traub, 1973, p. 210).

In a comparison of the three models (Hambleton and Traub, 1971) they noted that under simulated conditions in which the three parameter model was the appropriate one, i.e., significant guessing was present

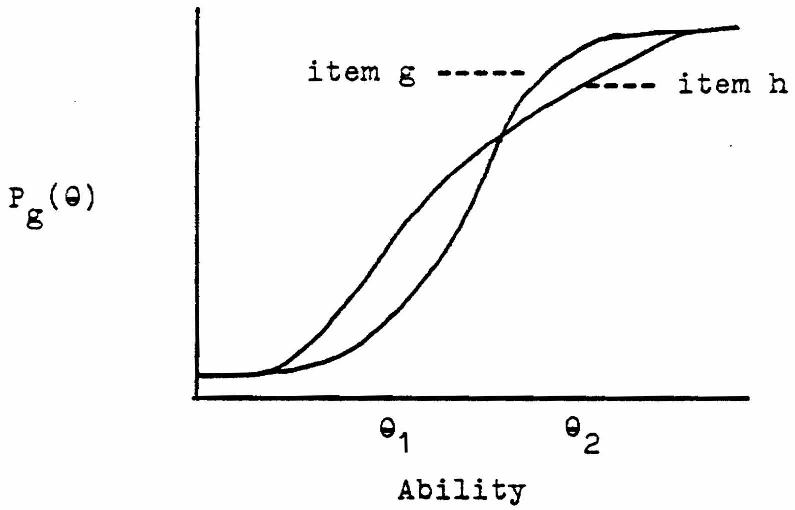


Figure 5

Item Characteristic Curves with
Different Discrimination Parameters

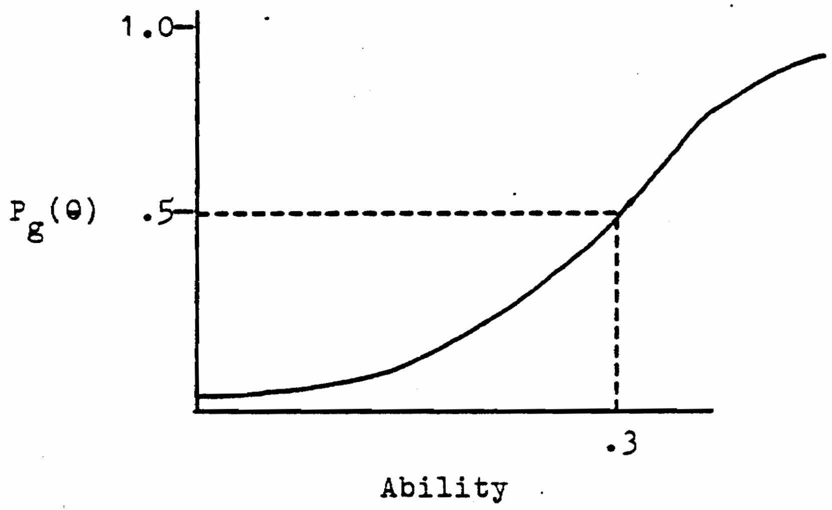


Figure 6

Example of a Rasch
Item Characteristic Curve

and items varied in discrimination, the one-parameter model was more appropriate than the two-parameter model at estimating the ability of low ability examinees. The Rasch model appeared to provide efficient estimates of ability across all ability levels when guessing was minimal, until the range of the distribution of discrimination parameters became large (.80 for that study).

In view of these estimation problems and the results of the comparison studies, the simpler Rasch model was selected as the basis for this analysis.

The Rasch model is based on an exponential function, $P_g(\Theta)$, relating the probability of success on item g for an examinee to his ability, Θ, on the trait being measured. Ability, Θ, is expressed on a continuous scale $(-\infty,\infty)$, but usually Θ is in the interval (-3,3). The item difficulty parameter b_g, is on the same scale as Θ; it is that point on the ability scale at which the probability of success is .5. In Figure 6, the difficulty of item g is .3.

Specifically, for the Rasch model the probability of success on item g for an examinee with ability Θ is given by:

$$P_g(\Theta) = \frac{\exp\ (\Theta-b_g)}{1+\exp\ (\Theta-b_g)} \qquad (1)$$

where b_g is the difficulty of item g.

The Rasch model permits the interpretation of item and ability parameter measurements as a ratio scale (Hambleton et al., 1977). For item g, a person of ability Θ_i has odds of success

$$O_{gi} = \frac{P_g\ (\Theta_i)}{1-P_g(\Theta_i)}$$

This may be simplified as

$$O_{gi} = \frac{\exp(\Theta_i)}{\exp(b_g)} \qquad (2)$$

Thus as Θ increases, the odds for success on a given item increases. As b_g increases, the odds for success for a given examinee decreases. Furthermore, if for examinees i and j $\Theta_i - \Theta_j = c$, for any real number c, then

$$O_{gi} = \exp(c)O_{gj}, \text{ for any item g.}$$

Also, if for items g and h, $b_g - b_h = c$, then

$$O_{gi} = \exp(-c)O_{hi}, \text{ for any examinee i.}$$

Again, consider equation (2) and take the natural log (ln) of both sides, giving

$$\ln(O_{gi}) = \ln[\exp(\Theta_i)] - \ln[\exp(b_g)] = \Theta_i - b_g.$$

Hence, the natural log odds of success is simply the difference of examinee ability and item difficulty.

Several procedures have been proposed for estimating the ability and item parameters for the Rasch model. The most widely used estimation procedure is the unconditional maximum likelihood solution described by Wright and Panchapakesan (1969). This procedure is used in the BICAL program developed by Wright and Mead (1978) and will be employed in this analysis.

In the Rasch model one ability estimate is obtained for each number right (raw) score (except scores of zero and perfect scores) and one difficulty parameter for each item. That is, for a test of 10 items there may be 9 ability estimates corresponding to raw scores 1-9 and 10 difficulty estimates for the items. A test provides no information on examinees of perfect or zero scores except that they have more or less ability than is measurable by the test; how much more or less is indeterminable.

Consider a test of N items. There will be N-1 ability estimates and N difficulty estimates. The situation may be depicted by the matrix in Figure 7.

There are Nx(N-1) possible probabilities depicted. For raw score group i we can observe the proportion who respond correctly to item g. The maximum likelihood estimation procedure provides a set of N difficulty estimates and N-1 ability estimates which, when functionally combined according to the Rasch model, best reproduce the observed frequency proportions. That is, the N+(N-1) estimated parameters make it "maximally likely" that the Nx(N-1) observed events occur (Ryan and Hamm, 1977).

The Rasch model assumes that the items are locally independent, i.e., the probability of an examinee answering an item correctly is not affected by his performance on other items in the test (Hambleton et al., 1977). Consequently, the joint probability of all responses depicted by the matrix in Figure 7 is the product of each of the separate probabilities.

The likelihood function then becomes

$$L=\prod_{g=1}^{N}\prod_{i=1}^{N-1} P_g(\theta_i) \qquad (3)$$

The first and second derivatives of L are taken with respect to the θ's and then with respect to the b_g's. The first derivatives are

POSSIBLE EXAMINEE BY ITEM INTERACTIONS FOR ESTIMATION PROCEDURE

		Item difficulties					
Raw scores	Abilities	b_i	b_2		b_g		b_N
1	Θ_1	p_{ii}	p_{12}	$\cdots$	p_{1g}	$\cdots$	p_{1N}
2	Θ_2	p_{21}	p_{22}	$\cdots$	p_{2g}	$\cdots$	p_{2N}
.	.	.	.		.		.
.	.	.	.		.		.
.	.	.	.		.		.
i	Θ_i	p_{i1}	p_{i2}	$\cdots$	p_{ig}	$\cdots$	$p_{i,N}$
.	.	.	.		.		.
.	.	.	.		.		.
.	.	.	.		.		.
N-1	Θ_{N-1}	$p_{N-1,1}$	$p_{N-1,2}$	$\cdots$	$p_{N-1,g}$	$\cdots$	$p_{N-1,N}$

FIGURE 7

set equal to zero and solved for parameter estimates. The resulting estimation equations are

$$a_g = \sum_{i=1}^{N-1} r_i P_g(\Theta_i) \qquad (4)$$

and

$$i = \sum_{g=1}^{N} P_g(\Theta_i) \qquad (5)$$

where

a_g = number of persons who get item g correct

i = a raw score

r_i = number of persons in raw score group i.

The two sets of equations (4) and (5) are solved to obtain parameter estimates (Wright and Panchapakesan, 1969).

The equations are solved by the Newton-Raphston iterative procedure. In this procedure, one starts with an initial estimate $x^{(0)}$ and obtains an improved estimate $x^{(1)}$. This new value is used as the starting estimate and the procedure is repeated until the estimates converge. If f(x)=0 is the equation to be solved, the value of x at the (n+1)th iteration is given by

$$x^{(n+1)} = x^{(n)} - (f(x)/f'(x))x = x(n)$$

where $x^{(n)}$ is the value of x at the n^{th} iteration (Wright and Panchapakesan, 1969).

For the Rasch model, the equations to be solved become

$$a_g - \sum_{i=1}^{N-1} r_i P_g(\Theta_i) = 0$$

and

$$i - \sum_{g=1}^{N} P_g (\Theta_i) = 0$$

The initial estimates for a sample of m persons are based upon the observed frequencies of success. The initial estimate of ability for persons in score group i is

$$\Theta_i^{(o)} = \ln [\frac{i}{N-i}]$$

and the initial estimate of difficulty for item g is

$$b_g^{(o)} = \ln [\frac{m-a_g}{a_g}]$$

where m is the sample size. The procedure is usually terminated when the average improvement in estimates is less than .001.

The item difficulties are centered around the mean item difficulty at each step so that a mean item difficulty of zero results. The asymptotic standard errors of the estimates are obtained from the second derivatives of the likelihood functions:

$$SE(b_g) = [\sum_{i=1}^{N-1} r_i P_g (\Theta_i)(1-P_g (\Theta_i))]^{-\frac{1}{2}}$$

and

$$SE(\Theta_i) = [\sum_{g=1}^{N} P_g (\Theta_i) (1-P_g (\Theta_i))]^{-\frac{1}{2}}$$

(Wright and Panchapakesan, 1969).

The procedure advocated by Wright and Panchapakesan (1969) for testing the goodness of fit of the Rasch model essentially involves examining the quantity a_{ig}, which represents the number of examinees in the i^{th} ability group (i.e., raw score group i) who responded correctly to item g. Then

$$y_{ig} = (a_{ig} - E(a_{ig}))/(V_{ar}(a_{ig}))^{\frac{1}{2}}$$

is distributed N(0,1). The quantity a_{ig} has a binomial distribution with parameters $P_g(\Theta_i)$ and r_i. Thus a χ^2 measure of the goodness of fit of the model can be defined as

$$\chi^2 = \sum_{\substack{i=1 \\ r_i \neq 0}}^{N-1} \sum_{g=1}^{N} y_{ig}^2 ,$$

which follows the χ^2 distribution with (N-1) (K-1) degrees of freedom, where N is the number of items and K is the number of non-zero score groups, when the model is appropriate.

Wright and Panchapakesan (1969) further defined goodness of fit tests for items as

$$\chi^2_g = \sum_{\substack{i=1 \\ r_i \neq 0}}^{N-1} y_{ig}^2 \qquad (6)$$

which is distributed as χ^2 with K-1 degrees of freedom.

Hambleton et al. (1977) pointed out several problems associated with the above χ^2 tests of fit. They noted that if any of the terms $E(a_{ig})$ is less than one, the deviates y_{ig} are not normally distributed, and a χ^2 distribution is obtained only by summing squared normal deviates. Furthermore, the test is sensitive to sample size, so that the null hypothesis that the model fits the data will always be rejected using the χ^2 test of fit for large samples. They do note that this is true for all statistical tests.

The general statistic for testing item fit in BICAL (Wright and Mead, 1978) is based on a modification of equation (6), in that residuals are summed over persons rather than score groups.

For examinee i with ability Θ_i for item g, let

$$x_{gi} = 1 \quad \text{if examinee i responds correctly to item g}$$
$$= 0 \quad \text{otherwise.}$$

Then for a sample of m persons on a test consisting of N items, a statistic for testing the fit of item g to the model is given by

$$X_g^2 = \sum_{i=1}^{m} Z_{ig}{}^2$$

where

$$Z_{ig} = \frac{X_{ig}-P_g(\Theta_i)}{\{P_g(\Theta_i)[1-P_g(\Theta_i)]\}^{\frac{1}{2}}}$$

$X_g{}^2$ is approximately distributed chi square with $\frac{(m-1)(N-1)}{N}$ degrees of freedom (Wright, Mead and Draba, 1976). The total fit statistic for item g listed by BICAL is

$$V_g = X_g{}^2 \cdot \left[\frac{N}{(m-1)(N-1)}\right] \qquad (8).$$

Wright and Mead (1978) state that V_g has an expected value of one and a variance of $\frac{2N}{(m-1)(N-1)}$.

Bias definitions.

Two basic definitions of item bias based on the objectivity and test of fit features of the Rasch model will be considered here.

The procedure for identifying bias proposed by Draba (1977) is based on the objectivity feature of the model. He stated that:

> ...difficulty estimates should be statistically equivalent for groups distinguished only by their ability distributions. If, on the contrary, difficulty estimates shift significantly from group to group, this suggests that the item interacts with particular characteristics of the groups such as race or sex (Draba, 1977. p. 1).

Difficulty estimates may not be identical, but they should be statistically equivalent for distinct groups. Statistical equivalence for estimates for white and black student groups will be determined by a t-test with

$$t_g = \frac{b_g(w) - b_g(b)}{\sqrt{SE_b(b_g)^2 + SE_w(b_g)^2}}$$

where $b_g(w)$ is the item difficulty estimate for item g for white students with associated standard error $SE_w(b_g)$, and $b_g(b)$, $SE_b(b_g)$ the corresponding estimates for black students. Draba warned that small standard errors resulting from large calibration sample size may result in trivial significant differences, and recommended adjusting the significance level in such situations (Draba, 1977). While multiple t-tests are required, one for each item of the test, it is assumed that the difficulty estimates are independent of one another.

Draba applied his technique to a 66 item reading test and considered white versus black, boy versus girl bias, with sample size n=900 for each group. Eleven items were identified as biased in the boy vs. girl analysis, and also eleven were identified in the black vs. white analysis. Content explanations were offered for many of these items.

The Durovic procedure, on the other hand, does not require that individual items fit the Rasch model and hence according to Durovic (1978) may be useful for a wider range of data sets. Durovic defined bias as follows:

> An item is biased for members of a group, if on that time, for members of the group, a mean square

> fit of the item to the Rasch model is obtained which differs, by greater than one, from the mean square fit obtained for members of the other group (Durovic, 1975, p. 4).

The mean square fit has an expected value of one, and "the observed data is expected to lie approximately one standard deviation from the expected data" (Durovic, 1978, p. 6). Durovic did not further explain the rationale for a difference of one in mean square fits as a criterion for item bias. He explained that the mean square fit indicates the degree to which a given item measures the same dimension as the other test items and that, hence, an item is biased if it does not relate to the dimension being measured in the same way for each group. He stated that a difference standard independent of the standard error, which is dependent on sample size, should be utilized. In a personal communication, Durovic stated that a difference of one was an arbitrary decision based on his experimental results using the fit statistics described by Wright and Panchapakesan (1969).

Durovic applied his procedure to 14 items used in selecting civil service personnel with n_b=367 and n_w=547. Two items were reported as biased, and the results were supported by a content analysis by black cultural experts (Durovic, 1975). Apparently no application on tests with greater than 14 items has been reported.

The rationale of a difference in mean square fit of one as a criterion for item bias needs to be further examined. The extent to which the mean square fit depends upon the number of items and the calibration sample size should also be investigated.

Durovic (1978) acknowledged the need for further theoretical work. He suggested that a ratio of mean square fits rather than a difference might be better, since such a ratio would be a ratio of independent chi-squares to which an F-test might be applied. Hence both differences and ratios of fit mean squares were considered in this study.

Both procedures were evaluated and direct comparisons were made. Furthermore, these procedures were compared with the traditional approach to defining item bias popularized by Cleary and Hilton (1968). The procedure for determining item by group interaction using traditional item statistics as a determination of bias as originally defined by Cleary and Hilton (1968) was undertaken using the method described by Angoff and Ford (1973). P-values (proportions of the sample answering correctly) for each item were calculated for each racial group. These p-values were transformed to normal deviates (z). The normal deviates were transformed to delta-values by $\Delta=4z+13$. A plot of the points (Δ_w,Δ_b) (one point for each item) was made. Angoff and Ford noted that the plot of these points normally forms an ellipse, and that items falling some distance from this plot may be regarded as indicating group by item interaction. They are items more difficult for one group than for the other, relative to the other items. Also the correlation coefficient represented by the deltas expresses the extent to which the items have the same rank order of difficulty in the two groups. Consequently, it inversely indicates group by item interaction (Angoff and Ford, 1973).

Data

The data set to be used is a sample from a 1974 state-wide testing program conducted at the eighth grade level in a southeastern state. The original testing program consisted of three mathematics subtests; three reading subtests; vocabulary (40 items), comprehension (20 items), everyday skills (45 items; a study skills subtest and several occupational/ attitudinal subtests. The three reading subtests and five selected items from the study skills subtest (110 items total) were initially chosen for this study. (Test data from 17,764 white students and 4,236 black students were available for this study. This represents approximately 1/3 of the total population who completed the four subtests.) In the interest of fiscal responsibility, initially a random sample of 6,108 of the 17,764 white students were selected, thus yielding n_w=6,108 and n_b=4,236 for the analyses. The raw data were dichotomously rescored with 1 correct, 0 incorrect.

The Rasch model, which was the basis for these analyses, assumes unidimensionality of test items. A principal components analysis was performed on the selected 110 items to determine the extent of the satisfaction of this assumption prior to application of the model. Separate principal components analyses were performed on both the black and white samples using procedure FACTOR of the Statistical Analysis System (SAS) (Barr et al., 1976) with 5 factors retained. There were 14 eigenvalues greater than 1; factor 1 for whites accounted for 18.4 percent of the variance, factor 2 accounted for 4.5 percent of the variance. For blacks, factor 1 accounted for 15.2 percent of the variance, factor 2 for 3.7 percent, and there were 22 eigenvalues greater than 1. Despite the relatively low percentage of variance accounted for by factor 1, this factor was dominant, with only 7 items having factor loadings less than .3 on this factor for the white sample. Two of the items were from the vocabulary subtests, three from the comprehension subtest, one from the everyday skills subtest. Based on content considerations and the mean scores for these items, the low loadings for five of the identified seven items were attributed to difficulty factors. The remaining two items were discarded from further analyses, resulting in 108 items.

Carroll (1945) showed that Pearsonian correlations between binary items measuring abilities tend to depend upon the difficulties of the items involved. He stated

> Factorial studies of items must be examined for the possibility that heterogenity of items in difficulty has given rise to Spurious factors (Carroll, 1945, p. 19).

McDonald and Ahlawat (1974) further attributed these spurious factors to the nonlinear regression of items or tests on the factor or factors of content. Specifically, they showed that binary variables that conform to the normal ogive latent trait model yields spurious factors due to departure from the linear model.

Consequently, it was conjectured that the wide range of item difficulties here and the hypothesized nonlinear regression of item scores on the latent ability were affecting the results of the principal components analysis. It was hence possible that the assumption of unidimensionality was justified for analytical purposes.

To further investigate this assumption, separate principal components analyses were performed for each of the three reading subtests for the white sample. Factor 1 for the vocabulary subtest accounted for 23.3 percent of the variance. The two items from this subtest with loadings less than .3 in the original analysis were also the only items with loadings of less than .3 on factor 1 for the subtest analysis. The results for the comprehension subtest were similar, with factor 1 accounting for 20.9 percent of the variance and the three previously identified items again having low loadings. Factor 1 for the everyday skills subtest accounted for 23.8 percent of the variance. Only one of the two items with low factor loadings for the original analysis had a factor loading of less than .3 on factor 1 for this subtest analysis; however, two additional items had relatively low factor loadings. The similarity of the results of the subtest analyses with the principal components analysis of the combined 108 items failed to indicate improvement in dimensionality by considering separate subtests.

However, preliminary item fit statistics did not support the assumption of unidimensionality. The item fit statistics produced by BICAL (Wright and Mead, 1978) based on calibrations of the selected 108 items for the white sample had a mean of .95 and ranged from .25 to 1.97; for the black sample the mean was 1.01, with a range of .58 to 1.58. The fit mean squares for both samples had an expected value of 1.0 and a standard error of .02.

The standard error of the item fit mean square is dependent upon sample size. To determine if the fit mean squares produced by BICAL were also influenced by sample size and to permit validation studies of the bias definitions as well, two random samples of approximately 400 each were selected from the white population (W1, W2) and two more samples of 400 each were selected from the black population (B1, B2). Wright and Mead (1978) noted that 400 persons are usually sufficient to calibrate items effectively; furthermore, Forster (1977, p. 13) noted that his analyses indicate that "no more than 200 persons were needed to obtain item calibrations virtually identical to those for the total population."

Attempts to refine the item set by eliminating misfitting items and recalibrating resulted in identifying more items as not fitting the model across all samples and frequently resulted in increased fit mean squares (see Table 2). Such findings are often attributable to multidimensionality of the item set (Hamm, 1977). Also, different items frequently were identified as misfitting both within the racial subsamples and between racial subsamples utilizing the four subsamples W1, W2, B1, B2 previously described.

Consequently, the original plan to combine the 108 items to form one conventional test was aborted. The three original subtests, vocabulary, comprehension, and everyday skills, were hence considered separately.

Satisfactory fits within at least one racial group were obtained by eliminating one item from the vocabulary subtest and two from the comprehension subtest, based both on the BICAL fit statistics and the results of the principal components analysis described previously. Results from the BICAL fit statistics and a content analysis indicate that the everyday skills subtest may not be unidimensional (see Table 3). All items were retained, but the possibility of multidimensionality for the everyday skills subtest must be considered in interpreting those results. The analyses reported were hence based on a 39 item vocabulary subtest, an 18 item comprehension subtest and a 45 item everyday skills subtest.

TABLE 2

DESCRIPTION OF BICAL[1] FIT STATISTICS COMBINING SUBTESTS

Sample	Fit Mn Sq Mean	Fit Mn Sq Range	SE[2]	No. of items
W1	.98	.33-1.71	.07	108
W2	.93	.19-1.78	.07	108
B1	.99	.67-1.42	.08	108
B2	1.00	.68-1.35	.08	108
W1	1.00	.34-2.15	.07	102
W2	.93	.19-1.45	.07	102
B1	.99	.66-1.37	.08	102
B2	1.00	.69-1.47	.08	102
W1	.97	.33-1.98	.07	95
W2	.95	.21-1.43	.07	95
B1	.99	.64-1.39	.08	99
B2	1.00	.58-1.38	.08	99

[1] BICAL (Wright and Mead, 1978)

[2] $SE = \left(\frac{2N}{(M-1)(N-1)}\right)^{\frac{1}{2}}$, N= no. items, m= calibration sample size

TABLE 3

DESCRIPTION OF BICAL[1] FIT STATISTICS FOR SUBTESTS

Vocabulary, N=39

Sample	Mean	Range	SE
W1	.99	.56-1.62	.07
W2	.98	.45-1.62	.07
B1	1.02	.82-1.20	.07
B2	1.02	.85-1.31	.07

Comprehension, N=18

Sample	Mean	Range	SE
W1	.95	.48-1.30	.07
W2	.97	.46-1.40	.07
B1	1.02	.16-1.33	.07
B2	1.03	.14-1.37	.07

Everyday Skills, N=45

Sample	Mean	Range	SE
W1	.88	.28-1.52	.07
W2	.86	.08-1.45	.07
B1	.98	.51-2.18	.07
B2	1.00	.55-1.88	.07

[1] BICAL (Wright and Mead, 1978)

Two-stage testing procedure

For accurate measurement, the difficulty level of a test should be appropriate to the examinee's ability level. With conventional tests this is possible only if the examinees are fairly homogeneous in ability (Lord, 1971). Testing procedures which involve varying test item difficulty according to the estimated ability of the examinee thus hold much promise for the accurate measurement of ability.

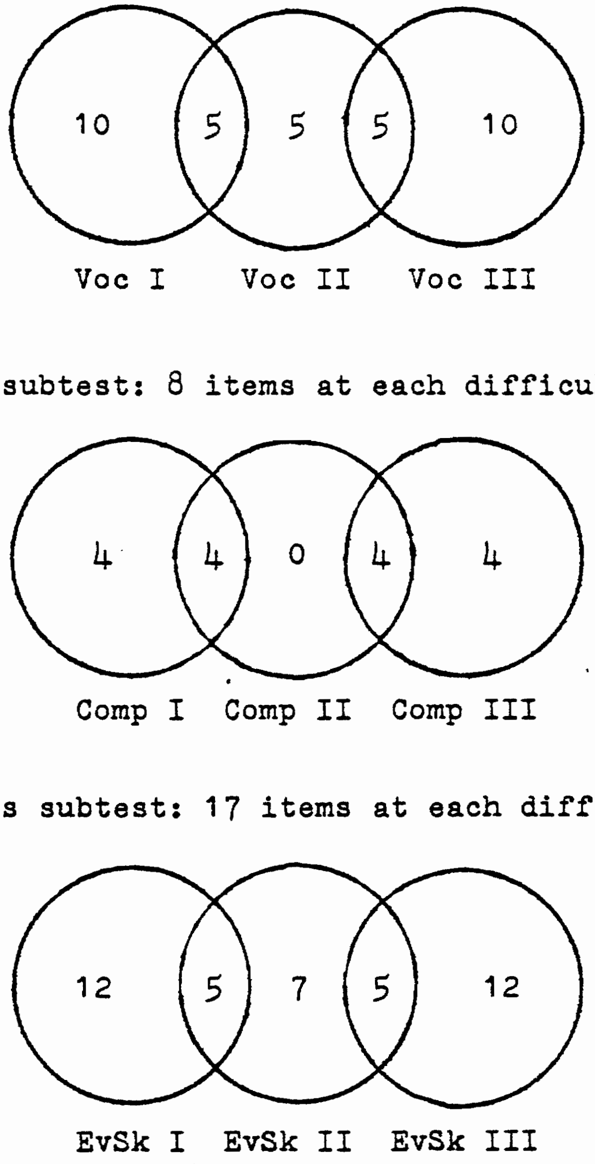
Two-stage testing is one approach to the implementation of such an "adaptive" testing procedure. The first stage consists of a short routing test which is used to obtain a rough estimate of the examinee's ability. Using this initial estimate the examinee is "routed" to a longer second-stage test consisting of items appropriate to his ability (Betz and Weiss, 1974).

In order to compare item bias on conventional testing procedures with two-stage testing procedures, a simulated two-stage testing procedure was developed based upon the real data set previously described. That is, one routing test and three second stage tests, each consisting of three subtests, were formed from the selected items

While it might seem reasonable to use a separate routing test for each of the three second stage tests, a single routing test was used which included representative items from each subtest. This decision was based on the assumption that a single overall ability estimate would be adequate and that it would be impractical to employ 3 separate routing tests in any real life testing situation. An initial rough estimate of ability for each examinee in both racial groups was made based on the routing test items alone. This initial estimate of ability was used to determine the set of second-stage test items used to obtain a final ability estimate for each examinee.

Rasch ability and difficulty parameters were first obtained separately for each subtest using the pool of items previously described. Each subtest was analyzed separately for bias. A routing test consisting of 10 items was formed from the pool of items fitting the model which had been determined to be bias-free according to the subtest results, based both on the fit statistics and on the equivalence of item difficulties. The routing test consisted of 4 vocabulary items, 2 comprehension, and 4 everyday skills items and, in accordance with the routing procedure advocated by Betz and Weiss (1974), consisted of items distributed uniformly in difficulty. The remaining 35 vocabulary items, 16 comprehension items, and 41 everyday skills items were used to form the subtests for the second-stage tests (Test I, Test II, Test III). Each second-stage test consisted of a 15 item vocabulary subtest, an 8 item comprehension subtest, and a 17 item everyday skills subtest. The subtests were formed from a linking network of items as indicated in Figure 8. Hence, a student routed to Test I was administered 3 subtests, Voc I, Comp I, EvSk I; a student routed to Test II was administered Voc II, Comp II, EvSk II, etc.

In accordance with the routing procedure advocated by Betz and Weiss (1974), Rasch ability estimates were obtained for each examinee based on the ten item routing test using BICAL. Examinees were then routed to that second-stage test (consisting of three subtests) with median difficulty most appropriate to his estimated



Vocabulary subtest: 15 items at each difficulty level

Comprehension subtest: 8 items at each difficulty level

Everyday skills subtest: 17 items at each difficulty level

Figure 8

Subtest Designs for Second-Stage Tests

ability level. The samples to be routed excluded those students in samples W1, W2, B1, B2. To ensure that the results for the comparison with conventional testing were not affected by disparity in sample size, random samples of approximately 400 students each were selected from both the black and white samples routed to each second-stage test.

Item and person parameters were then recalibrated for each of the three subtests within each of the three second-stage tests for the appropriate sample of examinees for each racial group. Then items within each subtests were examined for bias according to Draba's and Durovic's definitions. Hence items were examined for fit and difficulty equivalence between racial groups separately for each of the three subtests (Voc, Comp, EvSk) within each of the three second-stage tests (Test I, Test II, Test III). That is, the definitions were applied to the samples of black and white students routed to Test I for Voc I, Comp I, and EvSk I separately, and similarly for those routed to Test II and Test III. The combined results for each subtest were then compared with the results for that subtest based on the conventional testing procedure. Subsequent conclusions were essentially based on the results at the subtest level, comparing the individualized two-stage testing procedure with the conventional testing procedure for each subtest.

Chapter IV

RESULTS

The results of the calibration of Rasch item difficulty and ability parameters will be summarized first for each subtest of the conventional testing procedure, that is, for the vocabulary, the comprehension, and the everyday skills subtests. In a conventional testing procedure, recall that every item is administered to every student, regardless of the students' ability level. Bias analyses results based on the Angoff-Ford procedure, the Draba procedure, and the Durovic procedure will be reported for each of the three subtests for this testing procedure.

Following this, the bias analyses results for the three subtests will be summarized similarly for the two-stage testing procedure, in which there were three difficulty levels of each subtest. In a two-stage procedure, each student is administered only that level of each subtest appropriate to his ability level as determined by his performance on a routing test. The three procedures for the identification of bias will be considered at each difficulty level of each subtest.

Finally, the results of the conventional testing procedure will be contrasted with the results of the two-stage testing procedure to determine whether two-stage testing is effective in reducing the number of items identified as biased.

To facilitate discussion of the comparisons of the testing procedures, references will be made throughout to particular test items, e.g., V036 will refer to item number 36 in the vocabulary subtest. This will permit cross-referencing of biased items between the three reported bias identification procedures and between the two testing procedures. Since a content examination by a panel of educators with black cultural expertise failed to indicate substantive explanations for the bias identified statistically in this study (see Discussion, Chapter V), item content references would perhaps contribute little to the analyses. However, a few test items have been reproduced in Appendix D with permission of the publisher.

Analyses of bias for the conventional testing procedure.

Vocabulary subtest. The conventional 39 item vocabulary subtest tended to be inappropriate to the ability level of the target population for both racial groups, i.e., too easy for the white students, too difficult for the black students. Traditional item difficulties, i.e., proportion of sample responding correctly, tended to vary substantially between racial groups, reflecting the difference in ability distributions. Thus a bias definition which separates item difficulty estimates from the ability distributions of the samples is required for this situation.

The procedure for identifying item by group interaction suggested by Angoff and Ford (1973) for determining bias transforms the traditional item difficulties to normal deviates, by reference to a table of the normal curve, and then to delta values (Δ=4z+13). The delta values were plotted for the white sample by the black sample (see Appendix). Angoff and Ford noted that items falling at some distance from the plot may be regarded as contributing to the item by group interaction. Seven items (see Table 4) fell at some distance from the major axis of the ellipse formed by the delta values, and hence

TABLE 4

APPLICATION OF ANGOFF-FORD DEFINITION FOR CONVENTIONAL VOCABULARY SUBTEST

Conventional vocabulary subtest

Biased Items According to Angoff-Ford Procedure

VO 36

VO 1

VO 33

VO 19

VO 3

VO 38

VO 20

may be regarded as biased items. The correlation coefficient, .906, represented by the plot of the deltas for the two groups expresses inversely the item by group interaction. One problem associated with the use of the correlation coefficient as an indicator of item by-group interaction is its dependence upon test length. For example, the correlation coefficient may be reduced substantially more by a few biased items in a short test, but it is less likely to suffer the same consequence in a long test.

Draba's definition of bias is based upon statistical equivalence of the difficulty estimates between racial groups. T-tests were performed on the difficulty estimates both within and between racial groups to identify items biased by this definition and to determine the validity of Draba's definition. The results are summarized in Table 5, with a significance level of .01 used as a criterion as suggested by Draba (1977). Item difficulty estimates differed significantly between racial groups at the .01 level for 13 of the 39 items for W1xB1 and 14 for W2xB2, nine of which differed significantly between blacks and whites for both samples. Six of the seven items identified as biased by the Angoff-Ford procedure were also identified as biased according to Draba's definition. To determine the validity of this bias identification procedure, differences in item difficulties were also tested within racial groups. No item difficulties were significantly different at the .01 level between the two black samples. Two items resulted in significantly different item difficulties within the white samples. Hence there was apparently substantially more discrepancy in item difficulty estimates between

TABLE 5

APPLICATION OF DRABA'S DEFINITION FOR CONVENTIONAL VOCABULARY SUBTEST

Conventional vocabulary subtest

Significant t-tests (p=.01)

Groups	Item	Groups	Item
B1 x B2	*	W1 x W2	VO 28
			VO 32
W1 x B1	VO 2	W2 x B2	VO 2
	VO 3		VO 3
	VO 5		
	VO 6		VO 6
	VO 9		
	VO 10		
			VO 12
			VO 16
	VO 19		VO 19
	VO 20		VO 20
	VO 23		VO 23
			VO 26
			VO 28
	VO 30		
			VO 32
	VO 33		VO 33
	VO 36		VO 36
	VO 38		VO 38

* No significant differences

racial groups than within racial groups. Thus, Draba's definition apparently is detecting an item by racial group interaction, that is, possible bias.

In accordance with Durovic's definition of bias, the item fit statistics described by Wright and Panchapakesan (1969) (see Chapter III) were obtained for the 39 items for each of the four samples. These fit statistics are based on number-right (raw) score groups and are not the fit mean squares reported by BICAL. The BICAL fit mean squares are based on individuals and are estimated using six score groups. Hence a separate program was written to obtain the appropriate fit mean squares required by Durovic's definition of bias.

Durovic identified as biased those items whose fit mean squares differ by greater than one between racial groups. Three items are biased for W1xB1 based on Durovic's definition, and three for W2xB2, as summarixed in Table 6. No fit mean squares differed by greater than one within the white samples. However, three items resulted in fit mean square differences greater than one within the black sample. For this subtest, the largest difference in fit mean squares occurred between racial groups and was for the one item identified as biased for both W1xB1 and W2xB2. This is also the only item identified as biased by Angoff-Ford's procedure and by both Durovic's and Draba's procedures. Durovic suggested that applying an F-test to the ratio of the chi-square fit statistics might be more appropriate for the determination of bias. This resulted in three items having significant ratios at the .01 level between W1xB1, four items between W2xB2, no

TABLE 6

APPLICATION OF DUROVIC'S DEFINITION USING BOTH DIFFERENCES AND RATIOS FOR CONVENTIONAL VOCABULARY SUBTEST

Conventional vocabulary subtest

Differences in fit-mean squares greater than one			Significant ratios (p=.01) of fit mean squares
Groups	Item	Difference	Item
W1 x W2	*		**
B1 x B2	VO 7	1.335	VO 7
	VO 18	1.596	VO 11
	VO 37	-1.002	VO 18
			VO 37
W1 x B1	VO 18	-1.291	VO 34
	VO 36	2.034	VO 36
	VO 37	1.131	VO 37
W2 x B2	VO 7	1.030	VO 7
	VO 13	1.529	VO 13
	VO 36	2.717	VO 29
			VO 36

*no differences greater than 1.

**no significant ratios

items within the white samples, four items within the black samples. Neither differences nor ratios of fit mean squares satisfactorily distinguished results between and within racial groups. That is, there apparently is as much bias within as between racial groups according to this definition. Furthermore, results were not consistent for the two pairs (W1,B1), (W2,B2) of samples.

Residuals, as defined in Appendix A, were plotted by $(\Theta\text{-}b_g)$ as suggested by Wright, Mead, and Draba (1976) for one unbiased item and for two items identified as biased by at least one of the procedures. The residual plots (see Appendix) yielded little insight as to the nature of the bias for this subtest, perhaps due to the disproportionate number of students at different ability levels for each racial group. However, patterns indicating guessing trends, that is, large positive residuals at lower ability levels, were observed for both racial groups for the biased items.

Comprehension subtest. The results of the Angoff-Ford, the Draba, and the Durovic definitions of bias are considered below for the 18 item comprehension subtest. Preliminary to the consideration of Draba's and Durovic's definitions, based on the Rasch model, BICAL item difficulty estimates were obtained for the four samples, W1, W2, B1, B2. The comprehension subtest, similar to the vocabulary subtest, was below the mean ability level of the white sample, as indicated by negatively skewed ability distributions for both W1 and W2. The ability distributions of the black samples, however, indicated that the test was on target for this population, i.e., the mean Rasch ability estimate for each of the two black samples was approximately zero.

The delta value plots required for an examination of bias by the Angoff-Ford procedure have an associated correlation of .912. This may indicate less item by group interaction associated with this subtest than with the vocabulary subtest (r_{voc}=.906). This slight difference in correlation coefficients occurred despite the disparity in test length for the two subtests (n_{voc}=39, n_{comp}=18). Only one item, C08, fell at some distance from the plot and hence may be regarded as biased since it contributed to the item by group interaction.

Again, t-tests were performed both within and between racial groups to determine statistical equivalence of item difficulties as required for Draba's definition of bias. As in the vocabulary subtest a p-value of .01 was used as a criterion for determining bias. Table 7 summarizes the results. For W1xB1 five items resulted in statistically significant differences in difficulty estimates, two of which also resulted in significant differences between W2xB2. One item resulted in a statistically significant difference at the .01 level within the white samples; no items within the black samples, however, differed significantly in their difficulty estimates.

Durovic's definition of bias, based on the Rasch model as is Draba's requires the computation of differences in fit mean squares. A summary of these results is given in Table 8. Three items are biased for W1xB1 based on Durovic's definition; two items had differences in fit mean squares greater than one for W2xB2 and hence are biased for that sample pair according to Durovic's

TABLE 7

APPLICATION OF DRABA'S DEFINITION FOR CONVENTIONAL COMPREHENSION SUBTEST

Conventional comprehension subtest

Significant t-tests (p=.01)

Groups	Item	Groups	Item
B1 x B2	*	W1 x W2	CO 8
W1 x B1	CO 3 CO 6 CO 8 CO 9 CO 10	W2 x B2	CO 3 CO 8

*no significant differences

TABLE 8

APPLICATION OF DUROVIC'S DEFINITION USING BOTH DIFFERENCES AND RATIOS FOR CONVENTIONAL COMPREHENSION SUBTEST

Conventional comprehension subtest

Differences in fit mean squares greater than 1.0			Significant ratios (p=.01) of fit mean squares
Groups	Item	Difference	Item
W1 x W2	CO 7	-2.559	CO 8
	CO 10	-2.661	CO 10
	CO 19	-1.788	
B1 x B2	CO 7	-1.405	*
	CO 8	-1.997	
	CO 18	-1.146	
	CO 19	1.463	
W1 x B1	CO 9	1.586	*
	CO 10	-1.048	
	CO 18	1.029	
W2 x B2	CO 7	-1.085	*
	CO 10	1.144	

*no significant ratios

definition. Within the black samples, four items had differences in fit mean squares greater than one. However, the greatest differences in fit mean squares were found within the two white samples, with three items meeting Durovic's criterion of differences greater than one, two of which had differences in fit mean squares greater than two. The only significant differences in fit mean squares at the .01 level based on an F-test of ratios were two items within the white samples. The dissimilar results for the two between racial group pairs, each member of which was randomly selected from the same population, casts doubt on the validity of Durovic's procedure for the identification of item bias. This conclusion is supported further by results indicating substantial within racial group bias based on this definition.

Residual plots were again obtained for three items, one of which was unbiased. The residual plots showed a greater dispersion of points at the lower ability levels for whites for the biased items, which may reflect a greater guessing tendency among whites.

Everyday skills subtest. The markedly skewed ability distributions for W1 and W2 for this subtest indicate that the ability level of the white students far exceeded the difficulty level of this subtest. The ability distributions for the black samples were relatively platykurtic and somewhat negatively skewed, indicating that the test was perhaps too easy for this population as well, but that the black student group was more heterogeneous than the white with respect to the everyday skills measured by this test. Also recall that satisfactory fit was not obtained within either racial group for this subtest, indicating that either the assumptions of the latent trait models were not met or that another latent trait model might better describe these data.

The Angoff-Ford procedure resulted in a correlation of .954 for the item difficulty plot. This is the highest correlation among the three conventional subtests, perhaps indicating less bias associated with this subtest than with the vocabulary or comprehension subtests according to this definition. However, since this is also the longest subtest considered in these analyses, the high correlation coefficient may only be reflective of the dependence of this interaction indicator on test length. Items ES10 and ES16 were the only items identifiable as falling an appreciable distance from the narrow ellipse formed by the deltas.

Despite the evident misfit of the subtest fairly consistent Rasch item difficulties were obtained within both racial groups. The results of the t-tests required for Draba's definition of bias are summarized in Table 9. There were nine significant differences at the .01 level for W1xB1; there were six significant differences between W2xB2, five of which were also significant for W1xB1. However, there were no significant differences at the .01 level within either racial group. Both items identified as biased by the Angoff-Ford procedure are biased according to Draba's definition. Again, as with the vocabulary and comprehension subtests, Draba's procedure yielded fairly consistent results for the two between racial group sample pairs and identified substantially more items as biased between racial groups than within racial groups.

Differences and ratios of the number right (raw) score fit mean squares for the everyday skills subtest are summarized in Table 10.

TABLE 9

APPLICATION OF DRABA'S DEFINITION FOR CONVENTIONAL EVERYDAY SKILLS SUBTEST

Conventional everyday skills subtest

Significant t-test (p=.01)

Group	Item	Group	Item
W1 x W2	*	B1 x B2	*
W1 x B1	ES 7	W2 x B2	
	ES 10		
	ES 16		ES 16
	ES 21		ES 21
	ES 28		ES 28
	ES 38		
	ES 39		
			ES 40
	ES 41		ES 41
	ES 44		ES 44

*no significant differences

TABLE 10

APPLICATION OF DUROVIC'S DEFINITION USING BOTH DIFFERENCES AND RATIO FOR CONVENTIONAL EVERYDAY SKILLS SUBTEST

Conventional everyday skills subtest

Differences in fit mean squares greater than 1.0			Significant ratios (p=.01) of fit mean squares
Groups	Item	Differences	Item
W1 x W2	ES 44	1.387	*
	ES 45	1.183	
B1 x B2	ES 12	-1.104	ES 39
	ES 21	-1.008	ES 40
	ES 38	1.026	ES 41
	ES 39	1.869	
	ES 40	1.196	
	ES 41	2.587	
	ES 43	-2.620	
W1 x B1	ES 40	-1.046	ES 15
	ES 41	-2.861	ES 43
	ES 43	-1.067	
	ES 44	-1.844	
	ES 45	1.088	
W2 x B2	ES 42	-1.680	ES 42
	ES 43	-3.041	ES 43
	ES 44	-2.363	ES 44

*no significant ratios

Five items were identified as biased for W1xB1 based on Durovic's definition, three of which are the items identified as biased for W2xB2. There were two differences greater than two within the white samples. Note, however, that according to Durovic's definition, the greatest number of biased items occurred within the two black samples. Similar results were reported for the other two conventional subtests, raising doubts as to the validity of Durovic's definition for the determination of bias. For this subtest there were no significant ratios of fit mean squares at the .01 level within the white samples, three within the black samples, two for W1xB1 and three for W2xB2. Hence, as with the previous subtests, no apparent improvement in the validity of Durovic's procedure was obtained by considering ratios of fit mean squares rather than differences.

Analyses of bias for the two-stage testing procedure.

Students were routed to a second-stage test consisting of three subtests appropriate to their ability level as determined by their performance on a ten item routing test. The routing test was composed of four vocabulary items, two comprehension items, and four everyday skills items selected from the pool of unbiased items identified by the analyses for the conventional testing procedure.

Item difficulty and person ability estimates were obtained for the black and white samples, excluding those students in samples W1, W2, B1, B2. Hence, separate samples were used for the conventional and the two-stage testing procedures. Hence, separate samples were used for the conventional and two stage procedures. This may have

Five items were identified as biased for W1xB1 based on Durovic's definition, three of which are the items identified as biased for W2xB2. There were two differences greater than two within the white samples. Note, however, that according to Durovic's definition, the greatest number of biased items occurred within the two black samples. Similar results were reported for the other two conventional subtests, raising doubts as to the validity of Durovic's definition for the determination of bias. For this subtest there were no significant ratios of fit mean squares at the .01 level within the white samples, three within the black samples, two for W1xB1 and three for W2xB2. Hence, as with the previous subtests, no apparent improvement in the validity of Durovic's procedure was obtained by considering ratios of fit mean squares rather than differences.

Analyses of bias for the two-stage testing procedure.

Students were routed to a second-stage test consisting of three subtests appropriate to their ability level as determined by their performance on a ten item routing test. The routing test was composed of four vocabulary items, two comprehension items, and four everyday skills items selected from the pool of unbiased items identified by the analyses for the conventional testing procedure.

Item difficulty and person ability estimates were obtained for the black and white samples, excluding those students in samples W1, W2, B1, B2. Hence, separate samples were used for the conventional and the two-stage testing procedures. Hence, separate samples were used for the conventional and two stage procedures. This may have introduced slight differences in parameter estimates due to sampling, but it avoids problems of erroneously focusing on artifacts unique to the samples used for the conventional testing procedures. Routing was performed based on the Rasch ability estimates obtained. That is, students were selected for that second-stage test most appropriate to their ability level as determined by the Rasch ability estimate associated with their number-right (raw) score on the routing test. Based on characteristics of the second-stage tests, cut-off points were set based on number-right (raw) scores as described in Table 11.

Note the similarity in Rasch ability estimates between racial groups associated with the raw score cut-off points. Thus these cut-off points apparently routed students of similar ability in both racial groups to that second stage test most appropriate to their ability level. The size of samples routed to each second-stage test is disproportionate for the two racial groups, however. A similar difference in ability levels between the racial groups was observed for each subtest in the conventional testing procedure.

The BICAL fit mean squares for the items of the routing test ranged from .84-1.18 for the black samples, .75-1.09 for the white sample, indicating a unidimensional routing test. Random selections of approximately 400 examinees each were obtained from those white students routed to Tests II and III and from those black students routed to Tests I and II in order that all samples considered in these analyses were of similar size.

Vocabulary subtests. The ability distributions for each of the three 15-item second-stage vocabulary subtests indicated that effec-

TABLE 11

ROUTING TEST CUT-OFF POINTS

	Rasch ability	Estimate	Second-Stage Test	N	
Raw Score	Based on W	Based on B		W	B
0	-	-	I	11	82
1-3	-2.35 - -.85	-2.26 - -.85	I	397	1230
4-6	-.37 - -.49	-.38 - -.46	II	1932	1550
7-9	.93 - 2.22	.90 - 2.17	III	2615	524
10	-	-	III	710	41

tive routing did occur for this subtest. Recall that the 39-item conventional vocabulary subtests was "off-target" for both the black and white subpopulations, but the mean ability estimate for each racial group was approximately zero based on the appropriate level of the second-stage subtest. Also the BICAL fit mean squares indicated that the second-stage vocabulary subtests tended to fit the Rasch model better for the selected samples than did the conventional subtest. This may indicate that the second-stage subtests tended to be more unidimensional for their target populations than was the conventional vocabulary subtest.

In accordance with the Angoff-Ford (1973) procedure for determining bias, item difficulty delta-values were plotted for each of the three levels of the vocabulary subtest. Recall that a correlation coefficient of .906 was associated with the 39-item conventional vocabulary subtest; the second-stage subtests (15 items) resulted in r_{vocI}=.880, r_{vocII}=.763, r_{vocIII}=.617. The lower correlations for the second-stage subtests may be due more to reduced test length or to the restricted range of delta-values than to an actual increase in item by group interaction (bias). Apparent outliers, i.e., biased items, identifiable from the plots are listed in Table 12. Three of these seven items were identified as biased by this definition for the conventional testing procedure; the number of biased items is the same for the two testing procedures. However, as previously stated, these results may be reflective of the restricted range of values.

As required by Draba's definition of bias, t-tests were performed to determine the statistical significance (p=.01) of the differences

TABLE 12

APPLICATION OF ANGOFF-FORD DEFINITION FOR SECOND-STAGE VOCABULARY SUBTEST

Second-stage vocabulary subtest

Biased items according to Angoff-Ford Procedure

Subtest Level	Item
I	VO 2 VO 19
II	VO 3 VO 28 VO 29
III	VO 26 VO 36

TABLE 13

APPLICATION OF DRABA'S DEFINITION FOR SECOND-STAGE VOCABULARY SUBTEST

Second-stage vocabulary subtest

Significant t-tests (p=.01)

Subtest Level	Item
I	VO 11
II	VO 3 VO 28
III	VO 29 VO 33 VO 34 VO 36

in Rasch item difficulties obtained for the appropriate black and white samples for each level of the second-stage vocabulary subtest. Results are summarized in Table 13. A total of nine significant differences resulted, as compared with 13 for W1xB1 and 14 for W2xB2 for the conventional testing procedure. Four of the nine items identified as biased in the two-stage procedure were also identified as biased in the conventional testing procedure. Apparently stability in item difficulty estimates was improved by administering items only to student subpopulations of appropriate ability level.

Fit statistics based on raw score groups were obtained for each level, and differences and ratios of fit mean squares between groups computed. The results of the Durovic bias identification procedure are summarized in Table 14. The differences in fit mean squares apparently were substantially reduced by administering items only to the appropriate student subpopulations, as evidenced by the fact that only two differences in fit mean squares exceeded 1.0 for this subtest for the two-stage procedure.

Recall, however, that inconsistent results were frequently obtained for the conventional testing procedure when applying Durovic's definition.

Results from the two-stage testing procedure for the vocabulary subtest compared with those from the conventional testing procedure indicate both improved stability in item difficulty estimates and in fit mean squares between racial groups for the two-stage procedure. This apparent reduction in the number of biased items by two-stage testing may indicate that that which was originally identified as biased in the conventional testing procedure may in many cases simply

TABLE 14

APPLICATION OF DUROVIC'S DEFINITION USING BOTH DIFFERENCES AND RATIOS FOR SECOND-STAGE VOCABULARY SUBTEST

Second-stage vocabulary subtest'

Differences in fit mean squares greater than 1.0			Significant ratios (p=.01) of fit mean squares
Subtest Level	Item	Difference	Item
I	*		**
II	VO 3	1.001	**
III	VO 36	1.146	**

*no differences greater than 1

**no significant ratios

reflect artifacts associated with inappropriate difficulty levels in testing.

Comprehension subtest. The ability distributions for each of the three eight-item second-stage comprehension subtest levels indicated that the tests were generally appropriate to the estimated ability levels of most of the selected groups. The only exception was for level III for the white student sample; this high ability group had a negatively skewed ability distribution for the second-stage comprehension subtest (COMP III) to which they were routed, indicating that this subtest, although highest in difficulty level, nevertheless was still too easy for this student group.

The Angoff-Ford plots of the item difficulty deltas resulted in correlations of r_{compI}=.906, r_{compII}=.962, $r_{compIII}$=.821. The associated correlation for the conventional 18-item comprehension subtest was .912, indicating in one case a decrease in item by group interaction according to Angoff's and Ford's definition, despite a reduction in test length. The only obviously identifiable outlier, i.e., biased item, was C08 in subtest III. This item was the only item identified as biased by this procedure for the conventional subtest.

Application of Draba's definition of bias showed much improvement in the stability of item difficulty estimates for the comprehension subtest for the two-stage testing procedure compared with the conventional testing procedure as indicated in Table 15. Items are listed in ascending difficulty order based on calibrations for sample W1. Item difficulty parameters were calibrated within each

TABLE 15

ITEM DIFFICULTIES FOR COMPREHENSION SUBTEST

Item	Difficulty							
	Conventional test		Second-Stage Test					
			I		II		III	
	W1	B1	W	B	W	B	W	B
16	-1.31	-1.01	-.314	.011				
18	-1.24	-1.01	-.262	-.326				
1	-1.21	-1.26	-.678	-.583				
3	-.70	-.11	.452	.553				
4	-.68	-.64	.017	.034	-.758	-.866		
2	-.66	-.41	.042	.011	-.315	-.494		
11	-.38	-.67	.265	.022	-.792	-.905		
5	-.18	.08	.477	.279	-.058	.055		
20	-.14	.25			.032	.303	-1.068	-.758
12	.30	.46			.508	.337	-.749	-.486
6a	.46	.02			.365	.190	-.143	-.212
7	.59	.56						
17	.61	.95			1.019	1.381	-.129	-.070
13a	.63	.70						
8	.80	.14					.743	.069
9	.84	.31					.227	.166
19	1.35	.87					.786	.602

[a] These items were included in the routing test, and hence excluded from the second-stage tests.

NOTE. ___ Item difficulties are underlined for items idenfitied as biased by Draba's definition.

subtest, and since item difficulties within a subtest are adjusted to have a mean of zero, parameters differ between tests. Of interest here, however, are the differences in item difficulties within a subtest. For the items identified as biased in the conventional subtest, differences in item difficulties ranged from .44 to .66. Item difficulty estimates within the second-stage subtests were more consistent, however only one item (C08 in subtest III) resulted in a significant difference (p=.01) in item difficulties for the two racial samples. This item was identified as biased by the Angoff-Ford procedure as reported above and was also identified as biased for both interracial sample pairs in the conventional testing procedure. Note that three of the five items identified as biased for W1xB1 in the conventional testing procedure are among the most difficult items and hence may be inappropriate to the ability level of most examinees. When these items were administered only to high ability students (second-stage level III), this discrepancy in item difficulties was substantially reduced for two of these items. Observe that, contrary to what is usually assumed in biased testing situations, all three of these items were more difficult for white students than for black students.

Durovic's procedure (Table 16) does not indicate any improvement in the stability of the fit mean squares between racial groups by using a two-stage testing procedure. That is, more items have differences between racial groups in fit mean squares greater than one for the two-stage comprehension test than for the conventional subtest (see Table 8). No significant ratios of fit mean squares were reported for either procedure. Recall, however, that inconsis-

TABLE 16

APPLICATION OF DUROVIC'S DEFINITION USING BOTH DIFFERENCES AND RATIOS FOR SECOND-STAGE COMPREHENSION SUBTEST

Second-stage comprehension subtest

Differences in fit mean squares greater than 1.0			Significant ratios (p=.01) of fit mean squares
Subtest Level	Item	Difference	Item
I	CO 1	-1.215	*
	CO 5	1.013	
	CO 11	-2.391	
	CO 16	-1.198	
II	CO 4	-1.830	*
	CO 20	-1.573	
III	CO 8	-2.664	*
	CO 17	-2.072	

*no significant ratios

tent results were obtained based on Durovic's procedure within and between racial groups for the conventional comprehension subtest for both differences and ratios of fit mean squares.

The results of the analyses based on Draba's definition of item bias tend to support the conjecture that individualizing tests according to the ability level of the examinee will reduce the effects of cultural bias. That is, apparently much of what was masquerading as bias in the conventional testing procedure may be explained by factors associated with the mismatch of student ability and test difficulty. This conclusion is supported by results from both the vocabulary and comprehension subtests. Previous inconsistent results using Durovic's definition tend to minimize the effect of the results based on his procedure.

Everyday skills subtest. Unlike the ability distributions for the second-stage vocabulary and comprehension subtests, the ability distributions obtained for the three levels of the second-stage everyday skills subtest tended to be negatively skewed for both racial groups, but especially for the white samples. This indicates that although students were routed according to their ability level to one of three everyday skills subtests of varying difficulty, each subtest still remained too easy for its target population. (Recall that markedly skewed ability distributions were also obtained for the conventional everyday skills subtest). Subtest I corresponded to a rather platykurtic ability distribution for the black sample; however, all other distributions tended to be more peaked. Thus effective two-stage testing did not occur for this subtest. That is, the test items did not tend to be appropriate for the ability level of the students to which they were administered. However, considering the inappropriateness of the conventional everyday skills subtest, some improvement in tailoring the test for the target population may have been attained.

The correlation coefficient of .954 obtained by the Angoff-Ford procedure for the conventional everyday skills subtest exceeded those obtained for the second-stage everyday skills subtests, with r_{ESI}=.876, r_{ESII}=.636, r_{ESIII}=.890. The reduced correlation coefficients may be attributable more to the reduction in test length or to the restricted range of item difficulty deltas, however, than to an actual increase in item by difficulty interaction. The dispersion of the plots made it difficult to identify outliers for subtests I and II; one item, ES15, was an apparent outlier for subtest III and hence may be considered biased.

Table 17 shows that there are 13 items whose item difficulty estimates differed significantly at the .01 level between racial groups. These are the items which may be identified as biased according to Draba's definition. No reduction of bias has occurred with the two-stage procedure according to these results. This may be a result of the fact that the tests still were not individualized according to the students' ability levels.

Durovic's procedure for identifying bias (Table 18), like Draba's failed to indicate any reduction in bias for this subtest, although no significant ratios resulted in the two-stage procedure.

TABLE 17

APPLICATION OF DRABA'S DEFINITION FOR SECOND-STAGE EVERYDAY SKILLS SUBTEST

Second-stage everyday skills subtest

Significant t-tests (p=.01)

Subtest Level	Item
I	ES 2
	ES 7
	ES 10
	ES 24
II	ES 15
	ES 16
	ES 20
	ES 21
III	ES 15
	ES 28
	ES 34
	ES 39
	ES 40

TABLE 18

APPLICATIONS OF DUROVIC'S DEFINITION USING BOTH DIFFERENCES AND RATIOS FOR SECOND-STAGE EVERYDAY SKILLS SUBTEST

Second-stage everyday skills subtest.

Differences in fit mean squares greater than 1.0			Significant ratios (p=.01) of fit mean squares
Subtest Level	Item	Difference	Item
I	ES 5	-1.534	*
	ES 10	-2.164	
	ES 23	1.057	
II	ES 25	-1.101	*
	ES 27	1.171	
	ES 31	1.072	
III	ES 39	1.361	*
	ES 44	-1.999	

*no significant ratios

The results of the analyses for both the conventional and the two-stage testing procedures for the everyday skills subtest may be reflective of the fact that ineffective measurement of the acquisition of everyday skills has occurred. Routing did not result in examples of subtests individualized according to a student's ability level for this subtest, and, accordingly, results are not applicable toward tailored testing situations.

To investigate the effect of improved tailored testing on the stability of item difficulty estimates, level III of the everyday skills subtest was "readministered" to the middle ability groups of both racial groups, and, similarly, level II was "readministered" to the low ability groups.

The resulting ability distributions for both groups indicated that the "readministered" levels II and III were relatively appropriate for the new groups. Level II for the low ability groups apparently resulted in non-normal distributions. However, normality is not a prerequisite for the calibration of Rasch parameter estimates, and hence these ability distributions should not affect the results.

Draba's definition of bias was applied to the item difficulty estimates obtained (Table 19) for levels II and III. Note that there are two fewer items identified as biased for level II than for the original administration of level II. Hence improved tailoring of test to ability level of examinee did reduce the number of biased items. Since no appropriate ability group was available for the "readministration" of level I, results cannot conclusively be compared with those from the conventional testing procedure.

TABLE 19

APPLICATION OF DRABA'S DEFINITION FOR READMINISTRATION OF SECOND-STAGE EVERYDAY SKILLS SUBTEST

Readmininistration of second-stage everyday skills subtest

Significant t-tests (p=.01)

Subtest Level	Item
II	ES 20
	ES 24
III	ES 15
	ES 21
	ES 39
	ES 40
	ES 42

However, a trend toward the reduction of the number of items identified as biased with on-target testing is indicated.

Chapter V

SUMMARY, DISCUSSION AND CONCLUSIONS

Summary.

The purpose of this paper was to compare the effects on item bias of conventional testing procedures to an adaptive testing procedure. To adapt tests to the ability level of the student, tests were individualized according to ability level for each student by a two-stage testing procedure. A latent trait model was selected as the basis for this study to permit the separation of item characteristics from examinee ability, an essential ingredient for an objective definition of test bias. Specifically, the Rasch model was selected both because of the simplicity of the model and because bias research has been reported based on this model.

An actual conventional testing procedure and a simulated two-stage testing procedure, both based on subsamples of the same real data set, were used in this study. The test data considered were selected from three reading subtests: vocabulary, comprehension, and every day skills. Several items from each test were omitted from the vocabulary and comprehension subtests based on the results of a principal components analysis and tests of fit for the Rasch model, resulting in fairly unidimensional item sets. Efforts to improve the fit of the everyday skills subtest by systematically removing misfitting items were not successful. Thus the lack of unidimensionality may have adversely affected analytic results for this subtest.

Three bias detection procedures were used to facilitate the comparison of the effects on bias of the two-stage testing procedure. Two of the bias detection procedures, those formulated by Draba (1977) and Durovic (1975), are based on the Rasch model. Results were also obtained according to the Angoff-Ford (1973) procedure for the identification of test bias, which is based on classical test theory parameters. Two random subsamples (n=400) were drawn from each racial sample (resulting in samples labeled W1, W2, B1, B2) to validate the results obtained for each bias identification procedure based on the Rasch model.

The Draba (1977) definition of item bias is based on a comparison of the item difficulty estimates between racial groups. When statistical tests of equivalence were applied within racial groups for the conventional testing procedure, two items from the selected 39 from the vocabulary subtest resulted in statistically significant (p=.01) differences in item difficulty estimates within the white groups, as did one item for the 18 selected from the comprehension subtest. There were no statistically significant differences within the black groups for these two subtests and none within either racial group for the everyday skills subtest. The number of statistically significant differences identified within racial groups was quite small in all instances when compared with the number identified between racial groups by this procedure.

When this same procedure was used to verify Durovic's (1975) definition of bias satisfactory results were not obtained, however. Durovic defined as biased those items whose fit mean squares differ by greater than one between groups. Not only did more items result in differences in fit mean squares greater than one within racial groups than between racial groups, but the absolute values of the within group differences frequently exceeded the between group differences. Furthermore, unlike the results for the Draba procedure, consistent results were not obtained when considering the between racial group sample pairs W1xB1 and W2xB2 for Durovic's bias identification procedure. Considering ratios of fit mean squares as suggested by Durovic (1978) yielded little improvement over simple differences.

The Angoff-Ford (1973) procedure apparently identified more item by group interaction (i.e., bias) present in the comprehension and vocabulary subtests than in the everyday skills subtest. However, this may be the result of the differences in subtest lengths. Generally, more items were identified as biased based on the Rasch procedures than on the Angoff-Ford procedure. However, the Angoff-Ford procedure expresses item by group interaction, i.e., test bias, inversely in terms of a correlation coefficient and hence is designed more for detection of bias at the test level than at the item level, unlike the Durovic and the Draba procedures.

In order to compare the results of these bias identification procedures for the conventional test with a two-stage testing procedure, students were "routed" to one of three difficulty levels of each subtest based on their ability estimate on a ten item routing test formed from items previously identified as unbiased.

Individualizing tests to examinee ability level restricted the range of traditional item difficulties (proportion of sample responding correctly) at each level of the subtests. A restricted range of item difficulties may result in a reduced delta value correlation, a statistic required for the Angoff-Ford procedure. This restriction plus the reduction in test length may have affected the comparison of the results of the Angoff-Ford method for the detection of bias for the two testing procedures, making this method inapplicable for comparing bias trends for the two testing procedures. That is, it is not possible to determine if the differences in the magnitudes of the correlation coefficients were reflective of differences in item by group interaction or of restrictions in the item difficulty ranges for the three subtests or reductions in test lengths.

Durovic's procedure resulted in an increase in the number of items with differences in fit mean squares between racial groups of greater than one for the comprehension and the everyday skills subtests and a decrease for the vocabulary subtest. The inconsistencies of the results for this procedure in all analyses, both within and between racial groups, however, indicate that differences in fit mean squares cannot be justified as a basis for a definition of bias. Therefore, no conclusions will be made based on the comparison of applications of Durovic's definition for the conventional and two-stage testing procedures.

The Draba method for the identification of biased items resulted in a substantial decrease in the number of biased items for the two-stage testing procedure over the conventional testing procedure for the vocabulary and comprehension subtests. The number of items identified as biased for the everyday skills test remained approximately the same. However, an examination of the ability distributions indicated that effective tailored testing had not occurred for this subtest, i.e., the subtest difficulty levels were inappropriate (in this case too easy) for the groups to which they were administered. An attempt to improve test tailoring by "readministering" level II to the low ability group and level III to the middle abilility group indicated a trend toward the reduction in the number of biased items. This is in agreement with the results of the analyses for the other two subtests.

Discussion and conclusions.

It was proposed that individualizing tests according to the ability level of the examinee might have a positive effect in the direction of reducing the number of items that might be detected as biased. Such individualization of tests should reduce the effect of disturbances such as guessing or carelessness, possible explanations for identified bias.

Examination of item difficulty estimates indicates that a substantial number of the items identified as biased by Draba's procedure were at the extremes of the difficulty distributions. That is, many of those items identified as biased were precisely those items which were inappropriate to the ability level of the examinees in at least one of the racial groups and thus may have elicited guessing or carelessness tendencies. Analysis of residual patterns for a few items did indicate a greater tendency to guess on some items among the white sample. Such differential trends in guessing or carelessness may be one source of systematic discrepancies in item and test characteristics between racial groups. These trends are reduced by tailored testing strategies. This may account for some of the over-all reduction in the number of items identified as biased when two-stage testing was implemented.

Differential trends between racial groups for guessing or carelessness may not explain all items identified as biased, however. Identifying extraneous variables to which only one racial group has been exposed is a difficult and often highly subjective task.

A panel of four black educators with interest in the area of black cultural characteristics examined a sample of forty test items selected from those administered in order to determine those items which might be racially biased based on content considerations. Items identified as biased in the conventional testing procedure were embedded in a sample of unbiased items for each subtest. Draba's definition of bias was used as a criterion in selecting all items because of the consistency of results in applying this definition both within and between racial groups. They were told only the number of items in each subtest which had been identified as biased by statistical procedures. None of the items which had been statistically identified as biased were identified as biased unanimously by the panel. Only two of the items, C03 and ES16, were identified as biased by a majority of the panel. The panel's labeling of items as biased or unbiased appeared to be a chance occurrence.

This may indicate either the presence of very subtle content explanations for the bias or the absence of content explanations. The latter would tend to support the conjecture that much of what is statistically identified as biased may be systematic error attributable to inappropriate test difficulty levels.

The analyses considered here indicate a trend toward the reduction of the number of items identified as biased through a two-stage testing procedure. Apparently much of what was identified as bias may be explained by factors associated with inappropriate test difficulty level. More research is needed; not only are similar studies recommended based on two-stage testing, but also the effect on bias of other tailored testing strategies should be analyzed. While results are not conclusive, it is felt that this study indicates much promise for the reduction through tailored testing strategies of systematic measurement error which may be misinterpreted as actual cultural bias.

The racial groups considered here differed markedly in ability distributions. However, since latent trait models permit sample-invariant calibration of item parameters (Hambleton et al., 1977) this factor alone should not have resulted in the identification of biased items. Simulation studies based on test length and sample size similar to those examined here may be needed to further verify sample invariance for extreme differences in ability distributions. The two-stage testing procedure resulted in tests of reduced length and restricted item difficulties. The restriction in difficulties was of primary concern here. For the sake of completeness, however, stability of item difficulties should be compared under conditions of reduced test length alone.

A secondary purpose of this study was to examine and compare the Durovic and Draba definitions of item bias. Both definitions are based upon the objectivity of the Rasch model. The Draba procedure for identifying bias consists of testing the statistical equivalence of item difficulty estimates between cultural groups. Any significant shift in difficulty estimates is considered indicative of a group by item interaction or bias; that is, such an item may be interacting with some characteristic of the group besides ability (Draba, 1977).

The empirical results of these analyses support Draba's technique for the detection of item bias. This conclusion is based on the results of within and between racial groups analyses. Item difficulty estimates within racial groups were found to be quite stable; however, much more fluctuation was detected between racial groups. More research is needed on the effects of shifts in item bias on ability estimates. Draba (1977) considered the effects on ability of a test in which each item was equally biased. His results indicate "dramatic consequences for the classification and placement of students" (Draba, 1977, p. 10). Further studies are needed to determine the effect on ability estimates of tests, such as that considered in this study, for which only some items are differentially biased, that is, for which only some items demonstrate different item difficulties between racial groups.

The Durovic (1975) procedure for the identification of item bias consists of identifying as biased those items whose fit mean squares, based on number right (raw) score groups, differ by greater than one between racial groups. The empirical results of these within and between racial groups analyses do not support this technique as a valid method for the detection of bias. Since both the number of fit mean squares with differences greater than one and also the magnitude of the differences for the within racial groups analyses frequently exceeded those for the between racial groups analyses, it is felt that there is no evidence to justify any predetermined difference in fit mean squares as a criterion for identifying biased items. Also, these data showed no improvement by utilizing statistically significant ratios of fit mean squares as a criterion for identifying bias.

More research is needed on the relationship between fit statistics and sample characteristics, such as sample size and ability distributions of samples, as well as between fit statistics and test characteristics, such as test length and dimensionality. Perhaps therein will lie clues as to why Durovic's research provided empirical support for his bias identification procedure whereas the results reported here do not.

In conclusion, the analyses reported here indicate that appropriate matching of test difficulty and examinee ability level may reduce the number of items identifiable as biased. Hence, tailored testing strategies, such as the two-stage procedure considered here, may be instrumental in the crucial task of ensuring equal opportunity for all social groups by reducing differential cultural measurement error.

REFERENCES

Angoff, W.H. and Ford, S.F. Item-race interaction on a test of scholastic aptitude, Journal of Educational Measurement, 1973, 10, 95-106.

Angoff, W. H. and Huddleston, E. M. The multi-level experiment: a study of a two-level test system for the College Board Scholastic Aptitude Tests, (Statistical Report SR-58-21). Princeton, N.J.: Educational Testing Service, 1958.

Barr, A. J., Goodnight, J. H., Sall, J. P., and Helwig, J.T., A User's Guide to SAS 76. Raleigh, North Carolina: Sparks Press, 1976.

Betz, N. E. and Weiss, D.J., An empirical study of computer-administered two-stage ability testing (Research Report 73-4). University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973.

Betz, N. E., and Weiss, D. J., Simulation studies of two-stage ability testing (Research Report 74-4). University of Minnesota, Department of Psychology, Psychometric Methods Program, 1974.

Birnbaum, A., Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968.

Bowers, J., The comparison of GPA regression equations for regularly admitted and disadvantaged freshmen at the University of Illinois. Journal of Educational Measurement, 1970, 7, 219-225.

Breland, H. M., Stocking, M., Pinchak, B. M., and Abrams, M. The cross-cultural stability of mental test items: an investigation of response patterns for ten socio-cultural groups (Research Report PR-74-2). Princeton, N.J.: Educational Testing Service, February, 1974.

Burrill, L. E., Statistical evidence of potential bias in items and tests assessing current educational status. Paper presented at the Fourteenth Annual Southeastern Comference on Measurement in Education, December, 1975.

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

Carroll, J.B., The effect of difficulty and chance success on correlations between items or between tests. Psychometrika, 1945, 10, 1-19.

Cleary, T.A., Test bias: prediction of grades of Negro and White students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Cleary, T. A. and Hilton, T. L., An investigation of item bias. Educational and Psychological Measurement, 1968, 28, 61-75.

Cole, N.S., Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.

Darlington, R. B., Another look at 'cultural fairness'. Journal of Educational Measurement, 1971, 8, 71-82.

Draba, R. E., The identification and interpretation of item bias (Research Memorandum 26). Chicago: University of Chicago, Education Statistics Laboratory, March 1977.

Durovic, J. J., Test bias: an objective definition for test items. In Test bias: some conflicting unbiased views. Symposium presented at the Annual Convocation of the Northeastern Educational Research Association, October, 1975.

Durovic, J. J., Use of the Rasch model in assessing item bias. In What's happening in measurement: the use of Rasch and other latent trait models. Symposium presented at the meeting of the Eastern Educational Research Association, Williamsburg, Va., March 1978.

Echternacht, G., A quick method for determining test bias. Educational and Psychological Measurement, 1974, 34, 271-280.

Forster, F., Assessment and the Rasch model. Unpublished manuscript, 1977.

Forster, F., Everything you wanted to know about the Rasch model (but were afraid to ask). Unpublished manuscript, May 1977.

Hambleton, R. K., Swaminathan, H., Cook, L., Eignor, D. R., and Gifford, J. A., Developments in latent trait theory: a review of models, technical issues, and applications. Paper presented at the joint meeting of the National Council on Measurement in Education and the American Educational Research Association, New York, April 1977.

Hambleton, R. K. and Traub, R. E., Information curves and efficiency of the logistic test models. British Journal of Mathematical Statistical Psychology, 1971, 24, 273-281.

Hambleton, R. K., and Traub, R. E., Analysis of empirical data using two logistic latent trait models. Journal of Mathematical Statistical Psychology, 1973, 26, 195-211.

Hamm, D. W., Application of the Rasch model to classroom achievement test, standardized achievement test, and affective scale data. Paper presented at the annual meeting of the Southeastern Psychological Association, Hollywood, Florida, May 1977.

Linn, R. L., Rock, D. H., and Cleary, T. A., The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 24, 129-146.

Linn, R. L. and Werts, C. E., Considerations of test bias. Journal of Educational Measurement, Spring 1971, 8, 1-4.

Lord, F. M. and Novick, M. R., Statistical Theory of Mental Test Scores. Reading, MA.: Addison-Wesley Publishing Company, Inc., 1968.

Lord, F. M., A theoretical study of two-stage testing. Psychometrika, 1971, 36, 227-241.

Lord, F. M., A study of item bias using item characteristic curve theory. Paper presented at the Third International Association for Cross-Cultural Psychology Congress, Tilburg University, Tilburg, the Netherlands, 1976.

Lord, F. M., Practical applications of item characteristic curve theory. Journal of Educational Measurement, Summer 1977, 14, 117-138.

McDonald, R. P. and Ahlawat, K. S., Difficulty factors in binary data. British Journal of Mathematical Statistics and Psychology, 1974, 27, 82-99.

Mead, R., Assessing the fit of data to the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

Pine, S. M. and Weiss, D. J., Effects of item characteristics on test fairness (Research Report 76-5). University of Minnesota, Department of Psychology, 1976.

Potthoff, R. F., Statistical aspects of the problem of biases in psychological tests (No. 479). Chapel Hill, N.C.: Institute of Statistics Mimeo Series, Department of Statistics, University of North Carolina, August 1966.

Rasch, G., An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

Ryan, J. P. and Hamm, D. W., Estimating the parameters in the Rasch model. Paper presented at the annual meeting of the Southeastern Psychological Association, Hollywood, Florida, May 1977.

Scheuneman, J., Validating a procedure for assessing bias in test items in the absence of an outside criterion. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

Thorndike, R. L., Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Whitely, S. E., Models, meanings and misunderstandings: some issues in applying Rasch's theory. Journal of Educational Measurement, 1977, 14, 227-235.

Wright, B. D., Sample-free test calibration and person measurement. Paper presented at the Invitational Conference on Testing Problems, 1967.

Wright, B. D., Misunderstanding the Rasch model. Journal of Educational Measurement, 1977, 14, 219-225.(a)

Wright, B. D., Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-115. (b)

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

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

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

APPENDIX A

Let X_{gi} denote the item score for an individual with ability Θ_i on item g, i.e., $X_{gi}=1$ if the examinee responds correctly, 0 otherwise. According to the model,

$$E(X_{gi}) = P_g(\Theta_i) = \frac{\exp[\Theta_i - b_g]}{1+ \exp[\Theta_i - b_g]}$$

and

$$V_{ar}(X_{gi}) = P_g(\Theta_i).\ [1-P_g(\Theta_i)]$$

The residual, that is the difference between observed and expected outcome, becomes

$$x_{gi} = X_{gi} - P_g(\Theta_i)$$

Standardized, this residual becomes

$$Z_{gi} = \frac{X_{gi} - P_g(\Theta_i)}{\{P_g(\Theta_i).[1-P_g(\Theta_i)]\}^{\frac{1}{2}}}$$

if $X_{gi} = 1$,

$$Z_{gi} = \frac{1-P_g(\Theta_i)}{\{P_g(\Theta_i).[1-P_g(\Theta_i)]\}^{\frac{1}{2}}} = \exp\left(\frac{\Theta_i - b_g}{2}\right)$$

if $X_{gi} = 0$

$$Z_{gi} = \frac{0-P_g(\Theta_i)}{\{P_g(\Theta_i).[1-P_g(\Theta_i)]\}^{\frac{1}{2}}} =-\exp\left(\frac{\Theta_i - b_g}{2}\right)$$

APPENDIX B

Plots for the Angoff-Ford Procedure

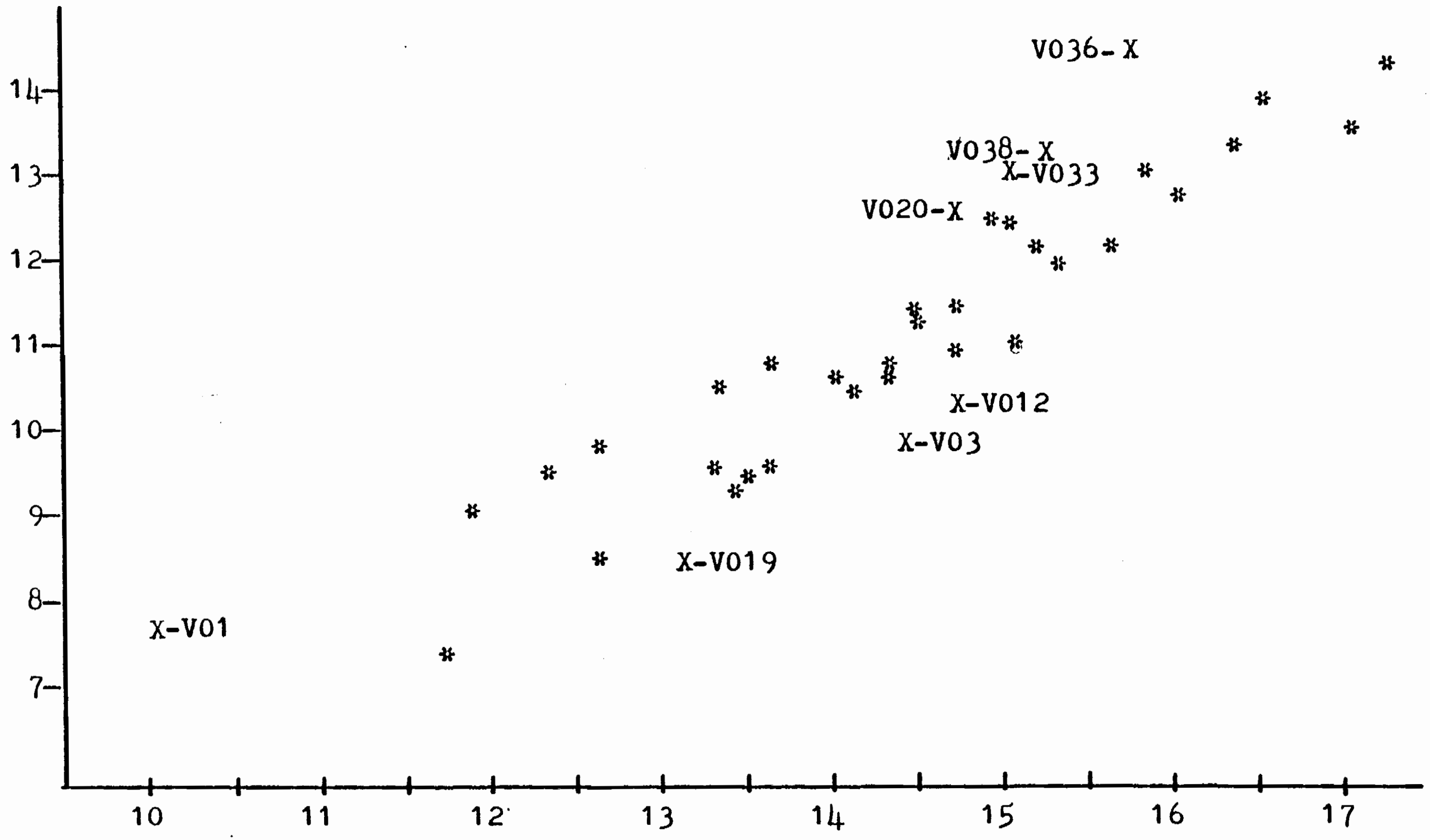


Figure 1

Plot of Deltas for Whites (down) by Deltas for Blacks
for Conventional Vocabulary Subtest

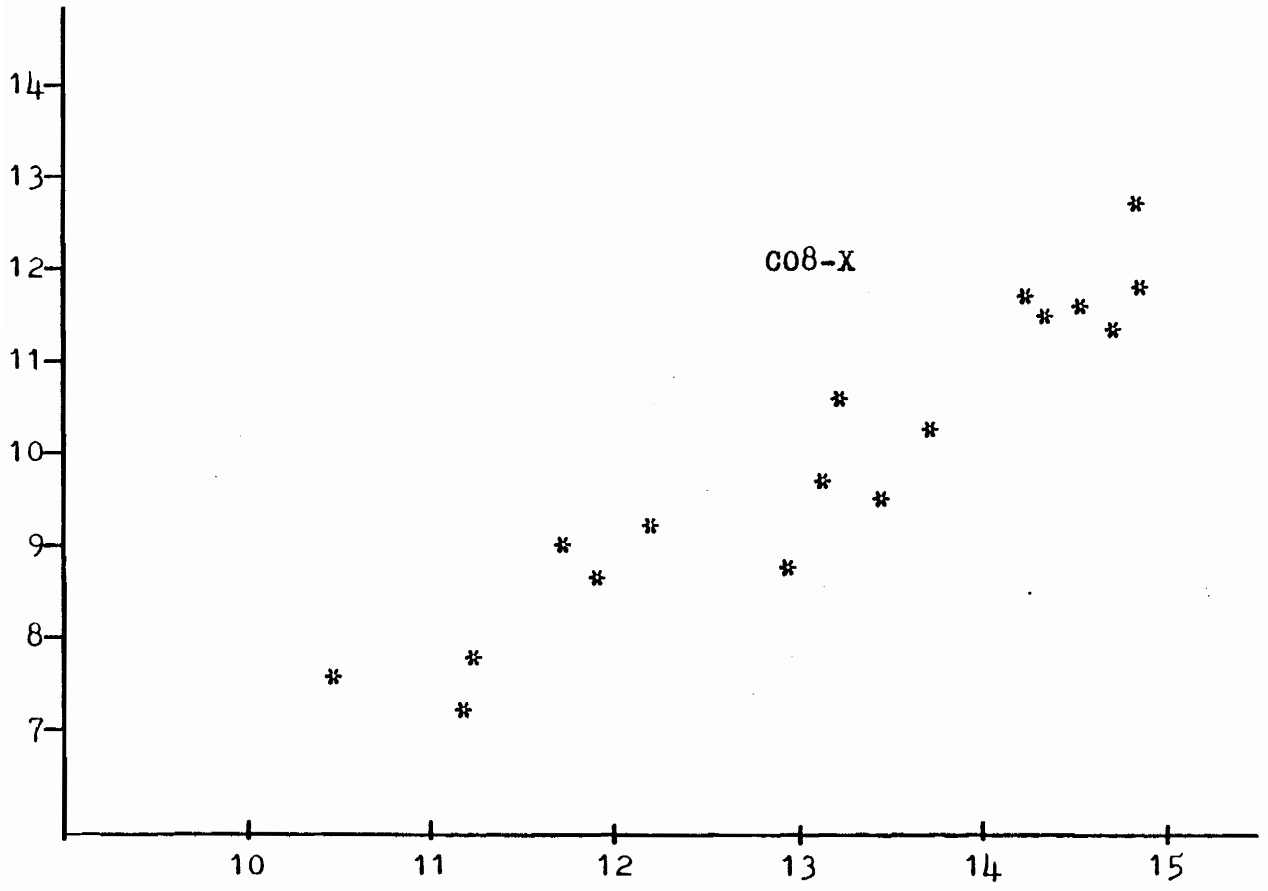


Figure ii

Plot of Deltas for Whites (down) by Deltas for Blacks
for Conventional Comprehension Subtest

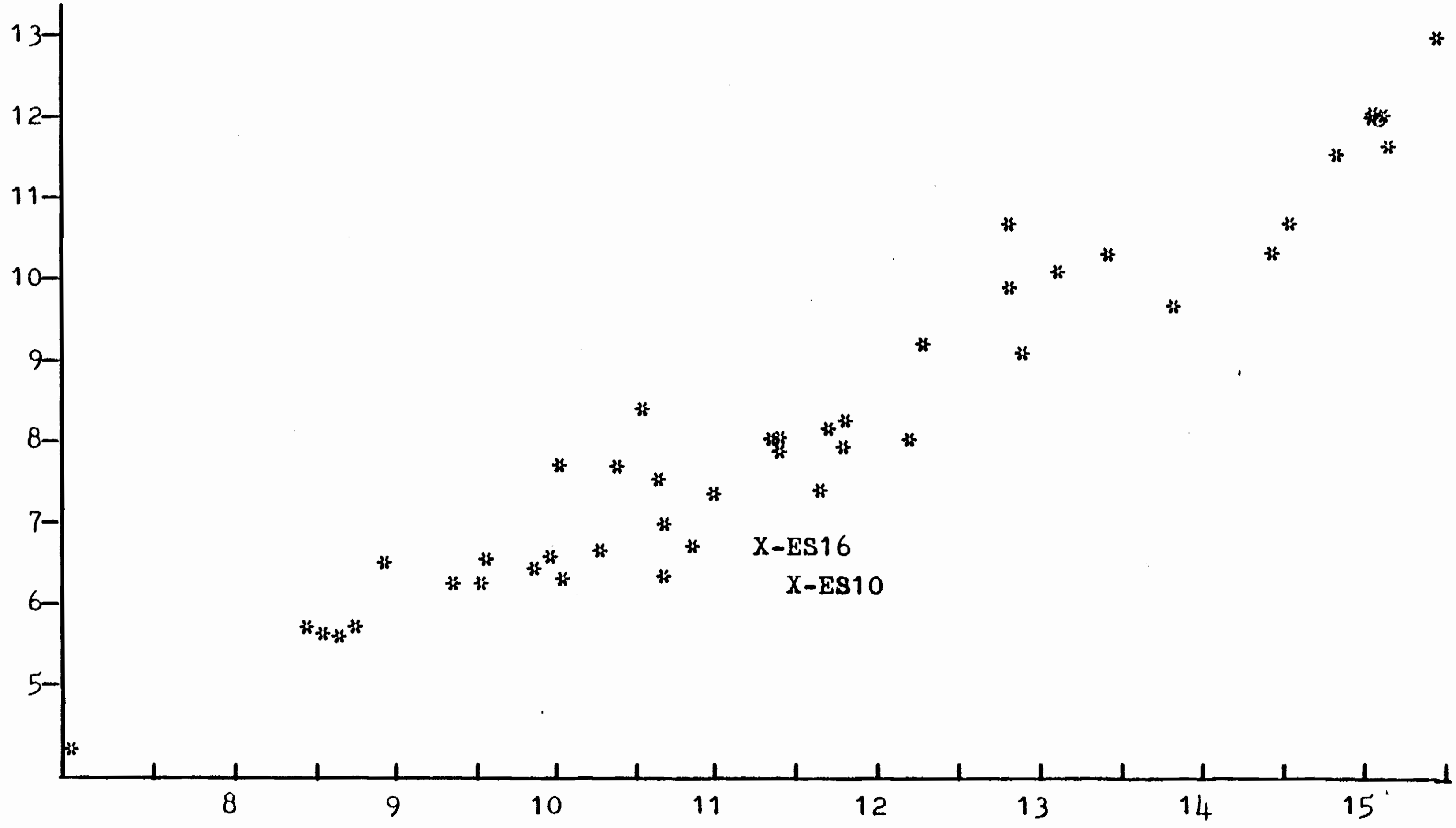


Figure iii

Plot of Deltas for Whites (down) by Deltas for Blacks
for Conventional Everyday Skills Subtest

APPENDIX C

Residual Plots

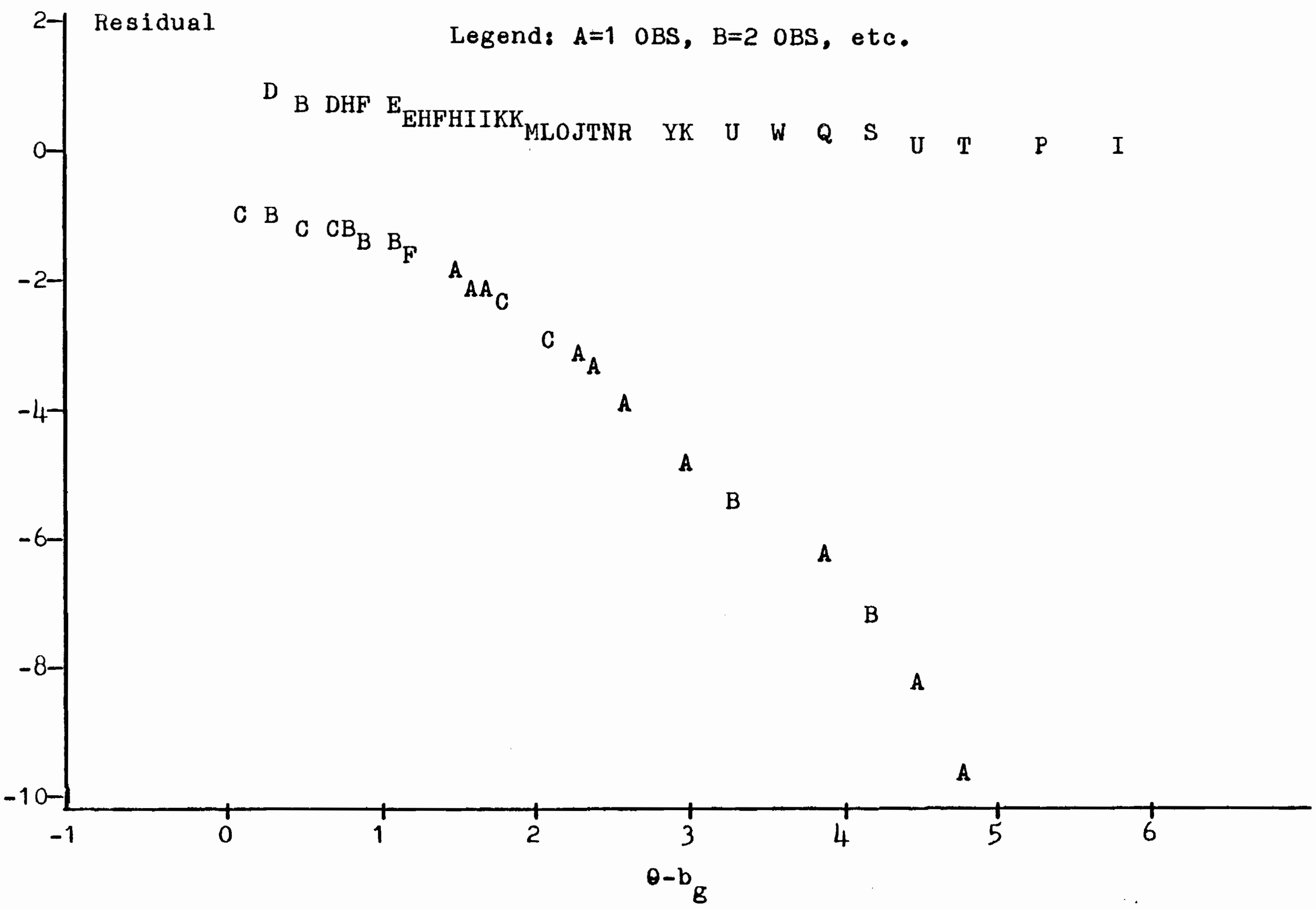


Figure 1

Residual Plot for V01 (unbiased) for White Sample

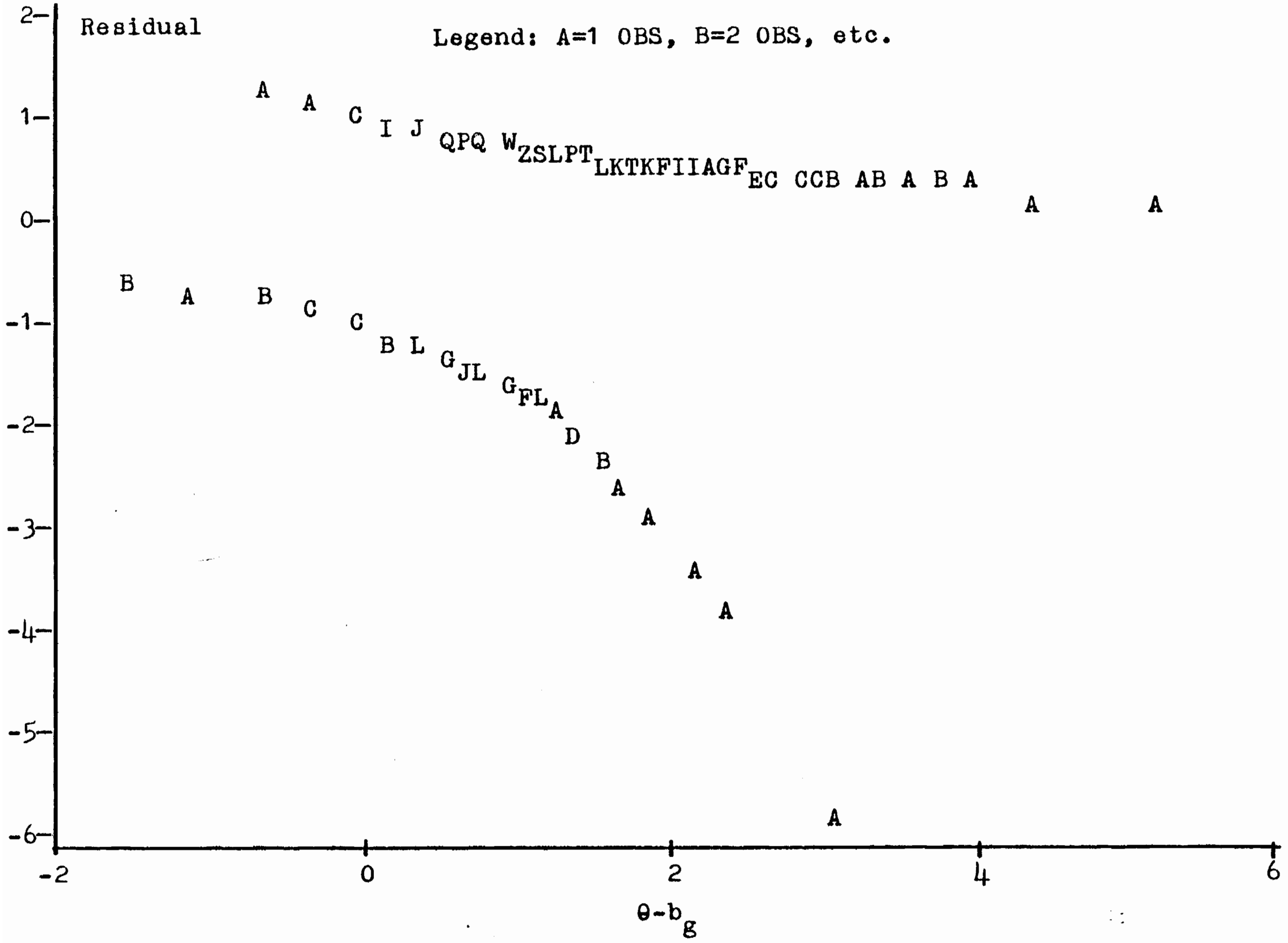


Figure ii

Residual Plot for V01 (unbiased) for Black Sample

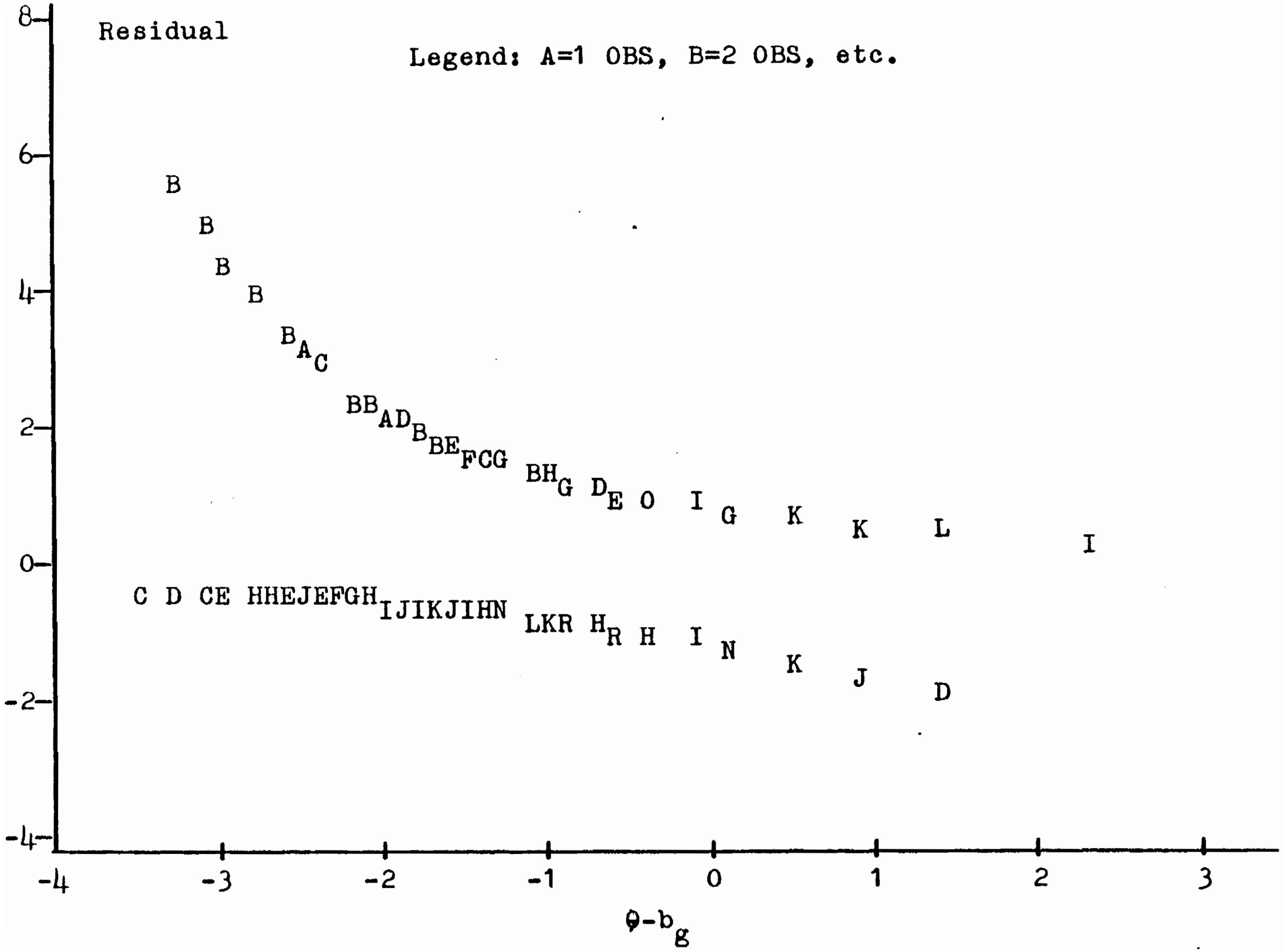


Figure iii

Residual Plot for V036 (biased) for White Sample

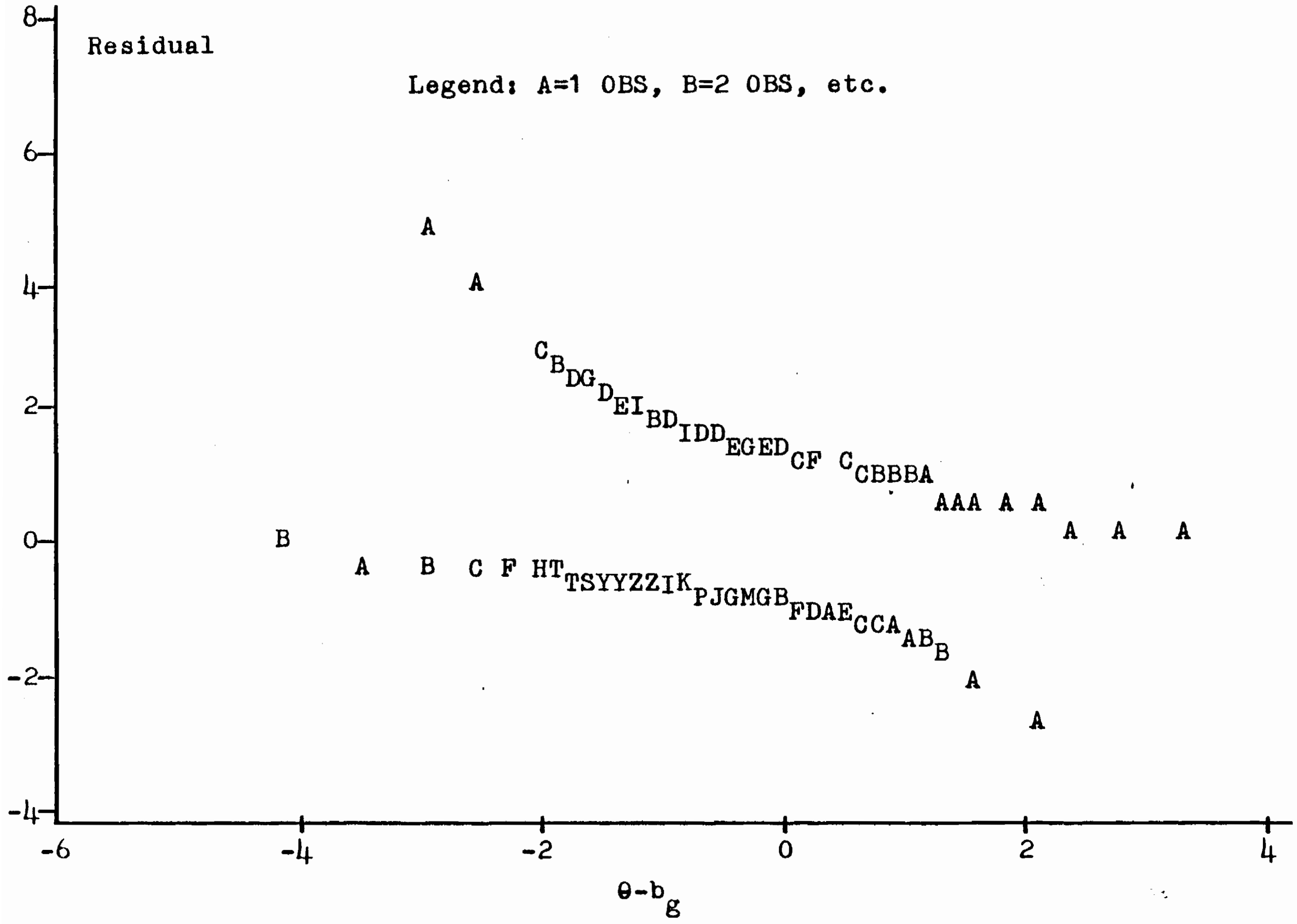


Figure iv

Residual Plot for V036 (biased) for Black Sample

APPENDIX D

A Sample of Test Items Statistically Identified as Biased

APPENDIX D

A Sample of Test Items Statistically Identified as Biased

I. Vocabulary subtest

23 logical question

A reasonable
B indispensable
C illegal
D offensive

33 advocate peace

A recommend
B prevent
C analyze
D oppose

36 facilitate the matter

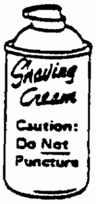
F expedite
G disturb
H copy
J discourage

38 preparatory step

F presumptuous
G preliminary
H preventive
J preferable

II. Everyday skills subtest

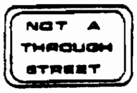
10 The caution on the can below means you should not



F shake the can
G make a hole in the can
H turn the can upside down
J put the can in the refrigerator

21 What does *tsp.* mean?

A foot
B pound
C teaspoon
D tablespoon

28 This sign means the street is 

F divided G one-way

H a dead end J being repaired

VITA

Carolyn Elizabeth Jones Lane was born in War, West Virginia on October 22, 1947. She graduated in 1969 from East Tennessee State University, Johnson City, Tennessee with the degree, Bachelor of Arts, in mathematics. She attended Virginia Polytechnic Institute from 1969-1972 under an NDEA fellowship, receiving the degree Master of Science in mathematics in 1971. Following a year of graduate school in mathematics and in education at the University of Virginia, she taught mathematics in the Virginia community college system. This teaching experience was followed by further graduate study at Virginia Polytechnic Institute and State University with a major in educational research.

Carolyn Elizabeth Jones Lane

Carolyn Elizabeth Jones Lane

A COMPARISON OF THE EFFECTS OF CONVENTIONAL TESTING AND TWO-STAGE TESTING PROCEDURES ON ITEM BIAS AS DEFINED BY THREE STATISTICAL TECHNIQUES

by

Carolyn Elizabeth Jones Lane

(ABSTRACT)

The purpose of the study was to compare the effects on item bias of conventional testing procedures to the effects of two-stage testing procedures. It is conjectured that much of the measurement error identified as bias can be explained by factors, such as guessing or carelessness, attributable to inappropriate matching of test difficulty level and examinee ability level.

Methods for detecting bias based on the traditional definition of item difficulty fail to separate test characteristics from the ability distribution of the respondent sample. The separation of item and ability parameters, however, is an essential ingredient for an objective definition of bias. Such objectivity in measurement is provided by the Rasch latent trait model, which consequently was selected as the basis for this study. Three definitions of bias were considered, two of which were based on the Rasch model.

The analyses were conducted using the scores of random subsamples (n=400 each) of black and white students on items selected from three reading subtests. The two-stage testing procedure was simulated using the real data set by "routing" students to one of three difficulty

levels of the subtests based on their Rasch ability estimates as determined by a ten item routing test. Results for the two-stage testing procedure were compared with those from the conventional testing procedure at the subtest level.

A reduction in the number of items identified as biased under conditions of appropriate matching of examinee ability levels and test difficulty levels was indicated by these analyses. Although the results are not conclusive, it is felt that individualizing according to the examinee's ability level offers promise in the direction of reading differential cultural measurement error.