1983

A Comparison of Two Methods for Detecting Test Item Bias

Elaine Kopera
Loyola University Chicago

A COMPARISON OF TWO METHODS FOR DETECTING TEST ITEM BIAS

by

Elaine Kopera

A Thesis Submitted to the Faculty of the Graduate School of Loyola University of Chicago in Partial Fulfillment of the Requirements for the Degree of

Master of Arts

January

1983

ACKNOWLEDGEMENTS

The author wishes to express her appreciation for the assistance given by Dr. Jack Kavanagh and Dr. Anne Juhasz.

A heartfelt thank you is also extended to the staff of the Joint Commission on National Dental Examinations, the Division of Educational Measurement and the Data Processing Department of the American Dental Association for their help and cooperation in completing this project.

Finally, a special thank you is extended to family and friends for their support, understanding and encouragement.

LIFE

The author, Elaine Barbara Kopera, is the daughter of Edward Matthew Weirich and Irene (Zbylut) Weirich. She was born November 10, 1949, in Chicago, Illinois.

Her elementary education was obtained in private schools in Arlington Heights, Illinois. Her secondary education was obtained in Sacred Heart of Mary High School, Rolling Meadows, Illinois where she graduated in 1967. In August, 1977, she received the degree of Bachelor of Arts with a major in psychology from Northeastern Illinois University, graduating with highest honors. While attending Northeastern Illinois University she was elected a member of Psi Chi in 1976.

TABLE OF CONTENTS

LIST OF TABLES

CONTENTS OF APPENDICES

CHAPTER I

INTRODUCTION

The field of testing has been the target of much criticism during the past decade. Specifically, a great deal of concern has been directed toward the issue of "fairness" in testing both at the test and the test item level (Cole, 1978; Shepard, 1980). In recognition of this concern a joint committee of the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education recommended, in 1979, that a complete revision of the 1974 Standards for Educational and Psychological Tests be undertaken which would include, among other topics, a provision for insuring test fairness (Cole, 1980). This action is an indication of the direction test developers and test users are following.

In view of the emphasis placed on the reduction of test bias by both the public and professional sectors, a need has arisen which calls for effective statistical procedures to be used in the detection of bias along with a set of guidelines to assist the user in choosing among them. This need has resulted in the development of a number of bias detecting methods and a number of research papers comparing their effectiveness (Cleary and Hilton, 1968; Lord, 1977; Ironson and Subkoviak, 1979).

For the purposes of this study procedures limited to the detection of bias at the item level will be examined. Test item bias is

defined as a difference in item response for test takers having the same ability but different group membership. It is the study of bias in the absence of any external criterion. Shepard (1980) refers to this type of bias detection as an empirical process rather than a logical analysis of test bias. Of the various item bias detecting procedures available there are four techniques which form the main core of those most frequently used and viewed as generating the most valuable results. Derivations and expansions of these four techniques are often employed (Rudner, Getson and Knight, 1980). The four major techniques are: the three parameter item characteristic curve (Lord, 1977); the chi square procedure (Scheuneman, 1979); the delta plot procedure (Angoff and Ford, 1973; Angoff, 1980) and; the Rasch Model (Wright, 1977). The advantages and limitations of each method are briefly summarized below.

The item characteristic curve has been found to be the most sensitive method for assessing item bias to date. It produces a sample-invariant item curve for each group which describes the probability of an examinee's answering an item correctly given his ability level. The strength of the method is its capacity to account for both the difficulty level and discriminative ability of the test item. Its weakness is that it is a highly complicated technique which is very expensive to compute. Also, it requires a sample size in excess of 1,000 and a test length of over 40 items in order to obtain accurate estimates of the three parameters.

Scheuneman's chi square procedure is a rough approximation of the item characteristic curve based on the proportions of correct

responses to an item by each group. Ability levels are clustered into intervals based on the total test score. According to Scheuneman (1980) an item is unbiased if the probability for success on an item is the same irrespective of the group to which the individual belongs. The advantage of using this technique rather than the item characteristic curve is its simplicity. Further, unlike the item characteristic curve, it has proven to be effective when using small samples or shorter tests (Scheuneman, 1979, 1980).

The delta plot procedure is the best alternative to the chi square and item characteristic curve (Shepard, Camilli and Averill, 1980). Bias is assessed as a measure of item by group interaction. The proportion of each group choosing the correct response is transformed to a normal deviate and then used to calculate a delta value. This is completed for each item. If an item does not have the same meaning or measure the same ability for the two groups the deltas will not fall in the same rank order. Deviate items indicate potential bias. A problem with the method is its inability to account for the discriminative ability of an item. However, this can be corrected for by matching groups for ability (Angoff, 1980). Like the chi square technique the delta plot can be performed with relative ease and is accurate using small samples, i.e., 50 or more, and short tests.

The Rasch model provides an item characteristic curve based on the assumptions that there is no guessing on the test and that all items are equal in their discrimination ability. With difficulty level as the only parameter of interest the curve indicates the

probability of a correct answer to an item over a range of examinee ability for each group. A major drawback to the technique is the difficulty in satisfying the assumptions. In general, it is not as effective as the previous three procedures in detecting test item bias. However, it is superior in other measurement application such as examinee ability.

Statement of the Problem and Rationale

As stated earlier in this paper, the need for guidelines to assist potential users in choosing among item bias detecting techniques has been demonstrated. Given the practical and theoretical limitations of the methods one must be able to select the method most appropriate to the situation. Although comparative studies have been conducted in the past, the need for this type of research continues (Rudner, Getson and Knight, 1980). Scheuneman (1980) states that most of the comparative studies pertaining to methods for assessing item bias have been completed using simulated data or large sample sizes. She notes a need for research using smaller samples and natural test data before the methods can be considered reliable across testing situations. According to Ironson and Subkoviak (1979), future studies are needed to "further evaluate the reliability, external validity and comparability of the methods." Also, there is a practical concern considering the range and variety of sample sizes required and cost associated with each method.

Purpose of the Study

The intent of this study is to contribute to the already

existing pool of knowledge pertaining to test item bias detection methodology by comparing the Scheuneman chi square procedure and the delta plot procedure. The test data used in the comparison will be taken from the biochemistry-physiology test of Part I of the July 1981 National Board dental examinations. The comparison is meant to determine the extent to which the two methods agree in their detection of biased items on this test. In the past, comparative studies involving these two methods have yielded a relatively high correlation of agreement between the two methods. This study will provide a reliability index of that correlation under a new testing situation. Given the replicative nature of this study it is hypothesized that there will be no significant difference in the level of agreement between the results of the two methods found in this study and that found in previous comparative studies.

CHAPTER II

REVIEW OF THE RELATED LITERATURE

The delta plot and chi square procedures are viewed as similar methods in that they both adjust their values for ability as measured by the total test score (Ironson and Subkoviak, 1979). They also both require unidimentionality of the test being examined (Shepard, Camilli and Averill, 1980). However, because each method has unique properties which necessitates individual attention, the procedures will be discussed separately followed by an overview of studies in which the two methods were compared.

Delta Plot Procedure

The delta plot methodology as described by Angoff and Ford (1973) is often referred to as the transformed item difficulty procedure (Ironson and Subkoviak, 1979; Shepard, Camilli and Averill, 1980; Rudner, Getson and Knight, 1980). It is derived from L.L. Thorndike's Method of Absolute Scaling (1925) in which the scores of tests developed for students at different grade levels were equated. This was accomplished by plotting identical items that were anchored in test forms used at different grade levels to determine any differences in the means and standard deviations for students in successive grades. In their 1973 paper, Angoff and Ford compare the performance of black and white candidates on the 1970 PSAT. They describe the delta plot as a measure of item by group interaction which is essentially a

graphical procedure. Angoff (1975), as well as Cleary and Hilton (1968) see item by group methods of bias detection as a measure of construct validity in which there is no clearly defined external criterion.

The delta plot method examines items to determine if particular items are more difficult for one group than for another. Accordingly, the rank order of item difficulty should not differ appreciably from one group to the next if the item is measuring the same ability for both groups. Angoff (1975, 1980) notes when items do not have the same meaning for the two groups under consideration, the deltas will not fall in precisely the same rank order for the two groups and the correlation between the two will fall below .98.

The procedure for computing the delta plot involves several steps. For each test item the proportion (p) of each group choosing the correct response is calculated. This p is then converted to a normal deviate, or z value, based on the standard normal curve transformation. A linear transformation of z values to delta values is then completed, expressed in terms of a scale with a mean of 13 and a standard deviation of 4 where: delta = 4z + 13. The advantage of the transformation to a delta value is the avoidance of a curvilinear plot of items, thus facilitating the analysis. After obtaining delta values a bivariate plot of the delta pairs for the two groups is prepared forming an ellipse extending from the lower left to the upper right of the graph. The scattergram of these pairs should hover around a straight line if the items are sensitive to the same ability for both groups. Any delta pairs significantly departing

from the line are considered representative of potentially biased items measuring something the other items are not. Sinnott (1980) found a perpendicular distance of .75 delta units or more to be an appropriate measure of significant distance in most instances. She did note the necessity of empirically validating the distance cutoff with each new set of data. The establishment of a baseline, e.g., comparing two white samples prior to a black versus white comparison, should be completed as part of the delta plot analysis. This is described by Angoff and Herring (1971) in their comparative study of the performance of Canadians and Americans on the LSAT.

A major criticism of the delta plot procedure is its inability to account for the discriminative ability of the item. As Cole (1978), Hunter (1975) and Lord (1977) have pointed out, unless all the items included in the plot have the same discriminative power, the delta plot may yield misleading results when the two groups under study score at widely different ability levels. Subkoviak, Mack and Ironson (1981) have also noted that in its earliest form the delta plot was found to be most sensitive in detecting group differences in item difficulty rather than group differences in item discrimination. This shortcoming was viewed by psychometricians as a principal reason for selecting alternate procedures considered to be more sensitive than the delta plot when feasible.

Stemming from this criticism, modifications of the delta plot intended to improve its statistical competence have been developed. Angoff (1980), Lord (1977) and Sinnott (1980) have found that matching groups on ability prior to the application of the delta plot procedure

serves as a reasonable solution to the omission of the discriminative power of items in the analysis. Another way to correct for this omission is to adjust the z scores by dividing the z corresponding to the p value for the item by the correlation between the item and the ability in question or by using the item-test biserial correlation (Angoff, 1980). Based on the new z values a delta plot can be prepared.

Another modification described by Angoff (1980) is concerned with the determination of the major axis line of the plot. He supports the evaluation of bias in individual items by measuring the distance from the major axis of the ellipse as defined by the items themselves rather than the 45° line, i.e., the equal difficulty, equal dispersion line. This should be done to reduce any chance of a true difference between groups being misinterpreted as item bias. This line is referred to as the "line of best fit".

Sinnott (1980) found that the likelihood of biased items contributing to the determination of the "line of best fit" is decreased if the following algorithm is used:

1) Calculate a preliminary line including all items.
2) Remove any items that are identified as biased based on their distance from the line.
3) Calculate a new line based on the remaining items.
4) Readmit the items removed in Step 2 to the set of points.
5) Repeat Step 2. If the items to be removed on the basis of the new line are the same as those removed the previous time the process is ended and the line calculated in Step 3

is considered the "line of best fit". If the items are not the same the process continues.

Finally, it should be mentioned that until recently the delta plot was considered limited to pairwise comparisons. Sinnott (1980) in her comprehensive paper on the delta plot procedure provides a technique applicable to more than two groups.

Scheuneman's Chi Square Procedure

The chi square method for detecting test item bias was first described by Scheuneman (1975) as an alternative to the highly sensitive yet highly impractical three parameter item characteristic curve used to estimate the probability that an item is biased. It was also developed to answer the need for a bias detecting method which would produce valid results using a small sample size while including indices of both item difficulty and item discrimination ability. Further, these indices are permitted to vary among items.

Scheuneman (1979) notes that this method is most accurately called a modified chi square technique. A typical chi square goodness of fit test is concerned only with the frequencies of persons in each category. In the case of the modified chi square the frequencies, i.e., the number of candidates from each group who score in each interval, is not of concern. Rather, the number of correct responses made by the candidates is of interest. The resulting effect is a loss of one degree of freedom for each ability group, when the sum of expected frequencies must equal the sum of the obtained frequencies. In addition, the degrees of freedom are restricted across ability

intervals where the sums of expected and obtained frequencies are most often not equal. Thus the degrees of freedom for the modified chi square procedure are (k-1)(r-1) where k is the number of groups and r is the number of interval groups formed.

Shepard, Camilli and Averill (1980) refer to the chi square method as a rough approximation of the latent trait model or item characteristic curve. Although it does not have the sample invariant properties of the latent trait model it does roughly equate group means on the total score distribution by creating matched score intervals. The probability of answering correctly should be the same in each interval irrespective of group membership.

The initial step in computing the chi square value is to establish discrete ability intervals on the total score scale of a homogeneous test. This must be done separately for each item and must meet certain criteria. There must be a minimum number of incorrect responses included in each interval assuring the probability of a correct response within each interval is less than one. The probability of a correct response within one interval is assumed to be constant. Also, there must be a minimum of 10-20 observed correct responses per cell to produce a valid outcome and, as usually recommended for chi square procedures, a minimum of 5 expected frequencies per cell. If these criteria are not adhered to an inflated chi square value may be obtained. There should be a minimum of 100 subjects to reduce the chance of error. Obviously the greater the sample size the greater the reliability will be. Depending on the number of candidates available, the total score range is divided into

3-5 ability intervals. Intervals should be selected so that the smallest cell frequency is about the same size for each ability level.

Once the ability intervals have been determined the total number of correct responses within each group and the sum of these totals is calculated. Also calculated is the total number of persons within each group scoring within each score interval and the sum of these totals. These data are then used to compute the chi square value with a large chi square signifying bias.

Shepard, Camilli and Averill (1980) note the chi square provides a measure of the degree of group differences because it is a test of significance based on probability rather than the arbitrary rules most of the other bias detecting methods are based on. However, Scheuneman (1979) suggests this may be a potential problem because the procedure does not include incorrect responses. As a result the modified chi square values may not approximate the chi square distribution. This error is most likely to occur when group sizes are appreciably different or cell frequencies are very large. A way of compensating for this, according to Scheuneman (1979) is to rank order the chi square values rather than using probability in defining bias. This can only be done where the same number of intervals has been used for each item.

Comparative Studies Including the Delta Plot and Chi Square Procedures

Once the need for effective methods of detecting test item bias had been met there still remains the question of which method to choose under which testing situation (Scheuneman, 1980; Ironson

and Subkoviak, 1979; Rudner, Getson and Knight, 1980, Rudner and Geston, 1982). Studies have been undertaken with the intention of answering that question by comparing the results of two or more bias detection techniques. In addition to examining the comparative performance of the techniques, results under a variety of test situations were considered. Among those studies were several which included both the delta plot and the chi square methods.

Subkoviak, Mack and Ironson (1981) compared the performance of the three parameter item characteristic curve (ICC3), the delta plot and the chi square techniques to determine which procedure is most sensitive to actual item bias. This was done by intentionally introducing items known to be biased into a sample of items taken directly from the College Qualification Test vocabulary test. The biased items were constructed similarly to those used on the Black Intelligence Test of Cultural Homogeniety. These items were inserted on a random basis within each block of five items on the test. The test was administered to a sample of black and white college students.

Of the three techniques examined the ICC3 was found to be most sensitive when bias was introduced into an actual testing situation. Of the two remaining methods the chi square technique was found superior to the delta plot. This was evidenced by the correlation coefficients between the amount of a prior bias originally built into the items and the amount of bias detected by the method. The authors recommended use of the delta plot when neither the ICC3 or the chi square method were practical. This recommendation is based on the computational simplicity and small sample requirements of the delta

plot.

Rudner, Getson and Knight (1980) investigated seven item bias detection techniques in a comparison using data produced with a Monte Carlo procedure. Both the amount and the type of item bias were specified a priori. The general findings of the investigation show the ICC3 and the chi square to correlate highest in their ability to detect generational bias. This was thought to be due to the similarity of the two methods both conceptually and empirically. As in the study by Subkoviak, et.al. the delta plot was seen as a suitable alternative to the ICC3 and the chi square because it is easily computed and provides a satisfactory estimate of item bias. Merz and Grossen (1979) also conducted a comparative study of six item bias detection procedures using data produced with a Monte Carlo procedure. Their results found the delta plot to have the highest agreement correlation with generated data followed by the ICC3 and the chi square.

In a comparison of six procedures for detecting test item bias, Shepard, Camilli and Averill (1980) found the delta plot and chi square methods the best alternatives to the sensitive but impractical ICC3. Specifically, correlations of agreement were highest between the delta plot and chi square, with chi square and ICC3 second and delta plot and ICC3 third.

The study was completed using the Lorge-Thorndike and Ravens Intelligence Tests. Comparisons were made between hispanic and white samples and black and white samples selected randomly from 4th, 5th and 6th grade students in one school district. In general, the

investigators found selection of method to be used in detecting item bias a major determinant of the amount of bias found. Consequently it is essential to know which method is the most appropriate for a particular test situation.

Another comparative study, conducted by Ironson and Subkoviak (1979), used data from the six subtest of the National Longitudinal Study (1972) taken by black and white 12th grade students. A white versus white baseline measure was established prior to the investigation. Also, groups were matched on ability to avoid the possibility that differences in ability would be mistaken for bias. Four detection procedures were examined. The findings of this study were quite similar to those of Shepard, et.al. (1980). The largest correlation of agreement was between the ICC3 and the chi square (.49), with chi square and delta plot second (.37), followed by delta plot and ICC3 (.24). The researchers point out that such an agreement should come as no surprise since all three methods attempt to control for the mean difference of the groups being compared and then concentrate on measuring the relative difficulty of the items.

The authors reference a supportive study by Rudner (1977b) comparing normal and hearing impaired subjects in which he also found the correlation of agreement between methods to have the same rank order, i.e., chi square and ICC3 (.67) first, chi square and delta plot (.59) second and ICC3 and delta plot (.31) third. Differences in the magnitude of correlation coefficients are attributed to Rudner's use of one subtest. Ironson and Subkoviak encountered considerable variability in their correlation coefficients across subtests with

some as high as .74, .50 and .37 respectively. Based on their comparison the authors consider the chi square and delta plot methods to be the most practical and efficient methods available.

In reviewing the various studies comparing available test item bias detection methods Scheuneman (1980) discovered that most of these studies used either simulated test data or very large sample sizes. She conducted a study using a small sample of black and white candidates taking two successively administered forms of the professional certification examination. She calculated the correlation of agreement between the delta plot, the chi square and the Rasch procedures. Her findings indicate that use of the chi square in combination with either of the other two methods provides more information on biased items than the other two methods combined or any of the three alone. Scheuneman contends this study supports all three methods as valid measures of item bias but also provides evidence for using combinations of procedures to produce a reliable and valid measure of bias. She encourages further research to assess the correlation of agreement between methods.

Rudner, Getson and Knight (1980) reviewed the psychometric rationale of six item bias detection techniques. In a discussion of the limitations and advantages of the approaches the authors conclude that each method has a specific strength making it useful under certain circumstances. Pertaining to the two methods this paper is concerned with the following was noted.

The delta plot has the limitation of being based on proportions that index the average difficulty of an item for a given group. As

a result only items illustrating differential group performance are identified. This weakness, however, can be minimized by using modifications such as matching on ability or adjusting z scores (Angoff, 1980), or modifying the calculation of the major axis (Sinnott, 1980). The chi square technique is limited in that the total score distribution of the two groups effect the expected cell frequency and may result in inflated chi square values when the observed scores are different. This inflation can be systematized when identical intervals are used for each item. This technique has the advantage of not being limited to detecting bias only when it favors one group because it examines the item for bias based on the distribution of correct responses across ability groups. The authors conclude that while neither of the two methods is capable of detecting all instances of test item bias, both do contribute to the information pool of test developers and test users.

After reviewing the comparative studies available it is apparent that the number of items found to be biased and the set of items found to be biased is dependent on the method of detection used. Because many of the studies have involved simulated data, large sample sizes or data taken from tests administered to candidates on the lower end of the educational spectrum, further research involving different test situations is advisable. Of the methods most commonly reviewed the chi square and delta plot methods are the two most consistently recommended for use. Because of this these two methods will be used in this research project under a new test situation involving a sample of graduate level students on a licensure

examination. Hopefully this will add to the pool of information available on test item bias research in general, while examining the potential extension of these two methodologies in particular.

CHAPTER III

METHODOLOGY

A study was conducted to determine the amount of agreement in test item bias detection between the delta plot procedure and the chi square procedure when used to compare the item responses of a sample of black dental school students to a sample of other dental school students. Specifically, the intent of the study was to determine the extent to which the two methods agree in their ability to detect item bias when used with data from the biochemistry-physiology test of Part I of the National Board dental examination. In addition to a comparison of the two item bias detection procedures, the study examined those items identified as biased by either or both methods to attempt a determination of the probable cause of bias toward either of the groups in the study.

Hypothesis

H_o: The delta plot procedure and the chi square procedure will detect a similar number and a similar set of biased items on the same examination.

Instrument

The National Board dental examinations are licensure examinations developed to assist dental licensing boards to assess whether a candidate possesses the necessary cognitive skills to practice dentistry. There are 11 separate examinations administered in two

batteries. Part I, which is typically taken after two years of dental school, consists of examinations on the basic biomedical sciences. Part II, which is typically taken during the final year of dental school, consists of seven examinations covering clinical subjects.

The data employed in this study were taken from the biochemistry-physiology examination of the Part I battery administered in July 1981. This examination consists of 100 multiple choice items selected by a test construction committee composed of subject matter experts. Descriptive test statistics for this examination include a mean of 64.58, a standard deviation of 12.41, a reliability coefficient (KR21) of .86, and 86 percent statistically satisfactory test items (F.B. Davis 27%). Examination results are reported in terms of converted scores with a mean of 85 and scores below 75 considered failures. A minimum raw score of 49 was needed to pass the July 1981 biochemistry-physiology examination.

Subjects

The sample consisted of 350 dental school students completing their second year of classes. These students were all taking Part I of the National Board examinations for the first time in July 1981. The total sample was divided into three equal groups, one group of black dental students and two groups of other dental students. The second group of other dental students was used to establish a baseline measure.

Group I contained 115 black candidates of both sexes chosen

from the 274 black students enrolled at the 60 U.S. accredited dental schools during the 1979 academic year. Of the 274 black students enrolled, 219 had been voluntarily identified by name in a demographic questionnaire previously administered by the Division of Educational Measurement of the American Dental Association. The names of these identified students were then matched with those candidates participating in Part I of the National Board dental examinations in July 1981. Those names appearing in both instances were selected as Group I subjects. This procedure identified 133 black candidates who took the Part I examinations in July 1981. The remaining 76 black dental students who had been identified by the Division of Educational Measurement questionnaire had apparently either left dental school or were not eligible for the Part I examinations in July 1981. There was no way to account for the 55 black dental students who chose not to be identified on the Division of Educational Measurement questionnaire. It was possible that some of these students participated in the July 1981 Part I examinations. However, considering these students were distributed throughout a total candidate pool of 4,314, it was unlikely their presence effected the investigation in any way.

After the black sample was identified the test score of each subject was recorded. To avoid any of the inherent problems of the delta plot procedure when used with groups unmatched for ability, Group II and Group III were selected by matching the test scores of the black candidates. This was accomplished through several steps to equate the groups for ability as much as possible.

Test scores for subjects in Group I were divided into five

score intervals: 71-75, 76-80, 81-85, 86-90, and 91-95. The dental school in which each black student was enrolled was identified. Within that school two additional students were selected by matching their scores with the score of the black student on the biochemistry-physiology examination. These two students were then randomly assigned to either Group II or Group III. A cross check was performed to insure that no subject was used more than once. In those cases when an exact test score could not be matched for either or both of the groups, a candidate whose score fell in the same score interval was selected. Matching test scores by interval may result in a small amount of variance but no more than would normally be attributed to a guess factor and therefore should not significantly effect the outcome of this study. In those cases when there was no match available from the same school, a school of comparable academic standing, as measured by the National Board annual quintile rating, which rank orders performance of dental schools on the National Board dental examinations, was selected and a matching subject was chosen from that school. Table 1 provides the number of subjects from each group falling within each score interval. Uneven groups resulted from some difficulty in matching on the computer. Because the first and the fifth intervals contain so few subjects they were combined with the second and fourth intervals respectively creating three ability intervals of similar size. This was done for purposes of data analysis.

Subjects with scores falling below 71 or above 95 were excluded from the study because of the difficulty in satisfactorily matching

Table 1

Number of Subjects From Each Group Within Each Score Interval

Groups	Score Intervals 71-75	76-80	81-85	86-90	91-95
I	7	33	46	20	9
II	7	33	46	23	9
III	7	33	46	22	9

candidates at these two extremes with students from the same or academically comparable dental schools. This resulted in an omission of 18 of the 133 black candidates who participated in the examination and accounts for the sample reduction of candidates in each of the three groups. All subjects in Groups II and III were selected randomly from those candidates meeting the requirements by which to match. The subjects in the three groups represent 31 of the 60 U.S. accredited dental schools and account for all geographical regions. Specifically, Group I has students from 23 of the 60 dental schools, Group II has students from 28 of the 60 dental schools, and Group III has students from 31 of the 60 dental schools. In Group I the majority of black students came from two of the 23 dental schools represented. This was due to black student enrollment being the highest at these two schools.

Statistical Analysis

Data from the biochemistry-physiology examination were analyzed using the two previously described methods of detecting test item

bias, delta plot and chi square, for the three groups.

Delta Plot Method. Sinnott's refinement of Angoff's delta plot procedure as previously outlined was applied. Each item had a delta value (Δ) for each of the three groups ($\Delta_1, \Delta_2, \Delta_3$). A bivariate plot of delta values was created for each comparison ($\Delta_1, \Delta_2; \Delta_1, \Delta_3; \Delta_2, \Delta_3$). Groups II and III were plotted against each other to provide a baseline for evaluating the procedure with these data. The computer program used to obtain the delta values and their plot is described in Appendix A.

The major axis line of the ellipse of points was obtained by calculating the line of best fit for each plot. The definition of the line of best fit and its determination, as taken from Sinnott, follows. It is the line that minimizes the perpendicular distances of the Δ-pairs and is the line that passes through $\bar{\Delta}^{M}, \bar{\Delta}^{M^1}$) and has the slope:

$$\frac{\sigma_{M^1}^2 - \sigma_M^2 + \sqrt{(\sigma_{M^1}^2 - \sigma_M^2) + 4r_{MM^1}^2 \sigma_M^2 \sigma_{M^1}^2}}{2 \sigma_M \sigma_{M^1} r_{MM^1}}$$

where M, M^1 are the item performance of the two comparison groups;

$\bar{\Delta}, \bar{\Delta}$ are the means of the item deltas for M and M ;

σ_M, σ_{M^1} are the standard deviations of the item deltas for M and M ;

and r_{MM^1} is the correlation between the item deltas of M and M .

After the preliminary line of best fit was computed the algorithm previously described to decrease the likelihood that biased items would contribute to the line determination was employed. This was completed using a fixed perpendicular distance $\geq$ .75 delta units as a cutoff. The computer programs used to obtain the line of best fit with this algorithm are offered in Appendices B through D.

Once the line of best fit was obtained the perpendicular distance (D_i) of each item (i) from that line was taken as a measure of item bias. Distance was computed with the formula:

$$D_i = \frac{AX_i + B\text{-}Y_i}{\sqrt{A^2 + 1}}$$

A distance of $\geq$ .75 delta units away from the line was considered representative of a biased item.

Chi Square Method. Scheuneman's chi square procedure was then applied to the data resulting in a 2 x 3 contingency table of correct item responses, corresponding with the two racial groups and three total score intervals. The chi square value was calculated comparing the observed number (O) of subjects who responded correctly to the item to the expected number (E) who would respond correctly if the likelihood of a correct response in that ability interval were the same for both groups. Each ability interval and each racial group contributed to the chi square value for each item with the entry $(O\text{-}E)^2/E$. This measure is summed over all ability intervals and racial groups to obtain the chi square index of bias for each item. Table 2 illustrates the computation of chi square for a single item.

A large chi square value indicates item bias. To determine

Table 2

Example of the Computation of X^2 for One Item

Total Raw Score on Test	Number with Scores in Each Range: Black B	Other O	Total T	Obtained Frequencies (No. with item correct): Black B_o	Other O_o	Total T_o	Proportion Correct (T_o/T) p	Expected Frequencies: Black (p.B) B_e	Other (p.O) O_e
46-59	30	30	60	27	23	50	5/6	25	25
60-68	46	46	92	36	37	73	73/92	36.5	36.5
69-83	39	42	81	29	34	63	63/81	30.33	32.67

$$\chi^2 = \sum \frac{(B_e - B_o)^2}{B_e} + \sum \frac{(O_e - O_o)^2}{O_e} = .446, \text{d.f.} = 2$$

whether the chi square value was significantly large, a chi square table was referenced at the .05 level of significance with (k-1)(r-1) degrees of freedom where k = the number of groups and r = the number of ability intervals. Because the same number of ability intervals was used for each item, obtained chi square values were also rank ordered as an alternate measure of item bias. The computer program used to calculate chi square values is provided in Appendix E.

Following the application of both procedures to the data the number of items and the set of items identified as bias by each of the methods was noted.

CHAPTER IV

RESULTS

Items Identified as Biased by the Delta Plot Procedure

The proportion (p) of each of the three groups choosing the correct response for each item and their corresponding delta values are listed in Table 3. The obtained delta values were used in calculating the line of best fit for each of the three comparisons. Application of Sinnott's line-fitting algorithm to the Group I, Group II comparison resulted in three iterations before determining the line of best fit for those data.

The preliminary line was calculated using the entire 100 test items. Employing a cutoff distance of .75 Δ units the 100 pairs were put through the line-fitting algorithm. The preliminary line had the slope

$$A = \frac{7.334 - 7.389 + \sqrt{(7.334 - 7.389)^2 + 4(.892)^2 7.334(7.389)}}{2(2.718)(2.708).892}$$

$$A = \frac{13.078}{14.721}$$

$$A = .888.$$

Thirty-nine items were removed during the first cycle of the algorithm because their distances exceeded .75 Δ units away from the preliminary line. The line was recalculated based on the remaining 61 items.

The second iteration of the algorithm resulted in a line with

Table 3

Proportion Answering Items Correctly and Corresponding Delta Values

Item	Group I		Group II		Group III	
	Proportion	Delta	Proportion	Delta	Proportion	Delta
1	0.98	21.44	0.95	19.55	0.92	18.70
2	0.95	19.50	0.91	18.29	0.93	18.95
3	0.82	16.62	0.81	16.56	0.85	17.08
4	0.77	16.00	0.83	16.82	0.86	17.23
5	0.84	17.04	0.89	17.90	0.86	17.38
6	0.98	21.44	0.98	21.49	0.97	20.29
7	0.90	18.03	0.91	18.29	0.92	18.48
8	0.79	16.24	0.86	17.25	0.89	17.88
9	0.68	14.85	0.84	16.96	0.82	16.54
10	0.85	17.18	0.88	17.73	0.86	17.38
11	0.90	18.23	0.86	17.40	0.93	18.95
12	0.86	17.34	0.82	16.69	0.82	16.67
13	0.77	15.89	0.84	16.96	0.87	17.54
14	0.82	16.62	0.85	17.10	0.78	16.06
15	0.66	14.66	0.81	16.44	0.80	16.29
16	0.81	16.49	0.77	15.97	0.70	15.10
17	0.74	15.56	0.74	15.54	0.69	15.01
18	0.79	16.24	0.80	16.32	0.79	16.17
19	0.90	18.03	0.83	16.83	0.80	16.29
20	0.73	15.46	0.82	16.69	0.75	15.72
21	0.84	16.89	0.80	16.32	0.78	16.06
22	0.75	15.67	0.81	16.44	0.85	17.08
23	0.89	17.84	0.86	17.40	0.94	19.22
24	0.55	13.48	0.59	13.86	0.71	15.21
25	0.88	17.67	0.76	15.86	0.80	16.29
26	0.73	15.46	0.86	17.25	0.87	17.54
27	0.91	18.44	0.92	18.72	0.94	19.23
28	0.65	14.57	0.59	13.86	0.61	14.08
29	0.49	12.87	0.53	13.34	0.60	14.00
30	0.83	16.76	0.94	19.24	0.88	17.71
31	0.69	14.95	0.58	13.77	0.67	14.72
32	0.97	20.77	0.90	18.09	0.95	19.53
33	0.45	12.52	0.64	14.39	0.77	15.95
34	0.71	15.25	0.69	14.94	0.71	15.21
35	0.81	16.49	0.52	13.17	0.55	13.47
36	0.86	17.34	0.64	14.48	0.74	15.51
37	0.70	15.15	0.58	13.77	0.46	12.61
38	0.72	15.35	0.80	16.32	0.80	16.42
39	0.42	12.17	0.37	11.70	0.47	12.70
40	0.75	15.67	0.70	15.14	0.72	15.31
41	0.56	13.57	0.58	13.77	0.64	14.45
42	0.77	15.89	0.75	15.65	0.58	13.82
43	0.50	13.04	0.53	13.26	0.50	12.96

Table 3 (continued)

Item	Group I Proportion	Group I Delta	Group II Proportion	Group II Delta	Group III Proportion	Group III Delta
44	0.50	12.97	0.38	11.79	0.45	12.53
45	0.48	12.78	0.39	11.88	0.45	12.53
46	0.64	14.47	0.57	13.68	0.46	12.61
47	0.60	14.01	0.40	11.88	0.49	12.87
48	0.47	12.70	0.52	13.17	0.50	12.96
49	0.31	11.05	0.48	12.75	0.46	12.61
50	0.37	11.71	0.31	11.06	0.41	12.09
51	0.28	10.65	0.36	11.61	0.36	11.56
52	0.70	15.05	0.58	13.86	0.62	14.26
53	0.23	9.99	0.28	10.67	0.34	11.37
54	0.48	12.78	0.69	14.94	0.65	14.54
55	0.64	14.39	0.61	14.12	0.56	13.56
56	0.65	14.57	0.60	14.03	0.62	14.17
57	0.39	11.90	0.37	11.70	0.29	10.79
58	0.30	10.85	0.25	10.25	0.29	10.79
59	0.38	11.81	0.43	12.32	0.44	12.44
60	0.75	15.67	0.74	15.54	0.71	15.21
61	0.50	13.04	0.63	14.30	0.77	15.95
62	0.52	13.22	0.63	14.30	0.53	13.30
63	0.50	12.96	0.63	14.30	0.65	14.54
64	0.28	10.65	0.36	11.61	0.43	12.27
65	0.63	14.29	0.43	12.32	0.48	12.79
66	0.70	15.15	0.82	16.69	0.80	16.52
67	0.74	15.56	0.75	15.75	0.75	15.73
68	0.22	9.88	0.19	9.56	0.27	10.49
69	0.55	13.48	0.73	15.44	0.68	14.82
70	0.42	12.17	0.33	11.25	0.33	11.28
71	0.48	12.78	0.35	11.43	0.30	10.89
72	0.81	16.49	0.84	16.96	0.86	17.23
73	0.74	15.56	0.70	15.14	0.62	14.17
74	0.28	10.54	0.31	10.96	0.33	11.28
75	0.84	17.04	0.62	14.21	0.72	15.31
76	0.77	16.01	0.80	16.32	0.80	16.29
77	0.50	13.04	0.53	13.26	0.47	12.70
78	0.62	14.20	0.64	14.48	0.62	14.17
79	0.49	12.87	0.45	12.49	0.41	12.09
80	0.53	13.31	0.48	12.75	0.49	12.87
81	0.56	13.57	0.56	13.60	0.54	13.39
82	0.43	12.26	0.28	10.67	0.25	10.28
83	0.56	13.57	0.57	13.68	0.59	13.91
84	0.53	13.31	0.45	12.49	0.49	12.87
85	0.58	13.83	0.32	11.15	0.34	11.37
86	0.31	11.05	0.33	11.25	0.39	11.92
87	0.57	13.75	0.37	11.70	0.36	11.56
88	0.44	12.34	0.38	11.79	0.49	12.87

Table 3 (continued)

Item	Group I Proportion	Group I Delta	Group II Proportion	Group II Delta	Group III Proportion	Group III Delta
89	0.45	12.52	0.48	12.83	0.50	13.04
90	0.62	14.20	0.62	14.21	0.60	14.00
91	0.32	11.15	0.53	13.34	0.56	13.56
92	0.94	19.19	0.98	21.49	0.99	22.54
93	0.80	16.37	0.73	15.44	0.72	15.31
94	0.37	11.62	0.36	11.61	0.39	11.92
95	0.95	19.50	0.98	21.49	0.89	17.88
96	0.96	19.85	0.91	18.29	0.89	17.88
97	0.06	6.81	0.08	7.28	0.06	6.78
98	0.84	16.89	0.76	15.86	0.68	14.82
99	0.20	9.63	0.43	12.32	0.39	11.92
100	0.80	16.37	0.80	16.32	0.85	17.08

a slope

$$A = \frac{6.880 - 7.830 + \sqrt{(6.880 - 7.830)^2 + 4(.979)^2 6.880(7.830}}{2(2.623)(2.798).979}$$

$$A = \frac{13.452}{14.370}$$

$$A = .936.$$

This calculation found 38 items beyond the .75 cutoff. A third line was calculated based on the remaining 62 items.

The third and final iteration produced a line with the slope

$$A = \frac{6.968 - 7.760 + \sqrt{(6.968 - 7.760)^2 + 4(.979)^2 7.760(6.968)}}{2(2.640)(2.786).979}$$

$$A = \frac{13.628}{14.401}$$

$$A = .946.$$

Again, using the cutoff distance of .75 Δ units it was found that the same 38 items were identified as those items exceeding the cutoff. This line was then taken to be the line of best fit. Table 4 provides a list of the 38 items identified as being potentially biased in the Group I, Group II comparison along with their respective distances from the line of best fit in delta units. Positive values indicate a bias in favor of Group I; the black sample. Negative values indicate a bias in favor of Group II; the same of other candidates.

The preliminary line for the Group I, Group II comparison had the slope

$$A = \frac{6.937 - 7.389 + \sqrt{(6.937 - 7.389)^2 + 4(.872)^2 7.389(6.937)}}{2(2.718)(2.634).872}$$

Table 4

Items Identified as Biased in the Group I, Group II Delta Plot Comparison with a .75 Cutoff

Item Number	Distance From the Line
1	1.02
5	-0.82
8	-0.88
9	-1.63
13	-0.92
15	-1.39
20	-1.02
25	1.20
26	-1.43
30	-1.98
31	0.76
32	1.62
33	-1.37
35	2.25
36	1.88
37	0.82
38	-0.82
44	0.82
47	1.48
49	-1.18
52	0.76
54	-1.59
61	-0.94
62	-0.82
63	-1.00
65	1.35
66	-1.23
69	-1.47
71	0.96
75	1.87
82	1.16
85	1.89
87	1.43
91	-1.55
92	1.94
95	-1.73
96	0.84
99	-1.84

$$A = \frac{12.042}{12.486}$$

$A = .964.$

Forty-four items were found to be beyond the .75 cutoff distance. The remaining 56 items were used in the second iteration of the algorithm which resulted in a line with the slope

$$A = \frac{7.042 = 7.651 + \sqrt{(7.042 = 7.651)^2 + 4(.977)^2 7.651(7.042)}}{2(2.766)(2.654).977}$$

$$A = \frac{13.747}{14.344}$$

$A = .958.$

Forty-five items were identifed as falling beyond the .75 cutoff using this line. The third and final iteration, which resulted in the identification of the same 45 items exceeding the .75 cutoff, produced the line of best fit with the slope

$$A = \frac{6.980 - 7.699 + \sqrt{(6.980 - 7.699)^2 + 4(.978)^2 6.980(7.699)}}{2(2.775)(2.642).978}$$

$$A = \frac{13.638}{14.341}$$

$A = .951.$

Table 5 lists the 45 items identified as potentially biased in the Group I, Group III comparison along with their respective distances from the line of best fit in delta units. Again, a positive value indicates a bias in favor of the black group while a negative value indicates a bias in favor of the group of all other candidates.

The Group II, Group III comparison required four iterations of

Table 5

Items Identified as Biased in the Group I, Group III Delta Plot Comparison With a .75 Cutoff

Item Number	Distance From the Line
1	1.72
4	-0.96
8	-1.27
9	-1.26
13	-1.27
15	-1.21
16	0.91
19	1.11
22	-1.07
23	-1.14
24	-1.24
25	0.86
26	-1.57
29	-0.78
30	-0.79
33	-2.44
35	2.10
36	1.20
37	1.72
38	-0.82
42	1.43
46	1.32
47	0.82
49	-1.03
53	-0.86
54	-1.23
57	0.87
61	-2.07
63	-1.11
64	-1.06
65	1.07
66	-0.97
69	-0.95
71	1.41
73	0.95
75	1.14
82	1.49
85	1.79
87	1.59
91	-1.65
92	-2.62
95	0.97
96	1.21
98	1.40
99	-1.50

the algorithm before determination of the line of best fit and resulted in the identification of 15 potentially biased items. The preliminary line had the slope

$$A = \frac{6.937 - 7.334 + \sqrt{(6.937 - 7.334)^2 + 4(.954)^2 6.937(7.334)}}{2(2.708()2.634).954}$$

$$A = \frac{13.218}{13.610}$$

$$A = .971.$$

After checking distances from the line 16 items were removed. The second iteration yielded a line with the slope

$$A = \frac{5.716 - 6.170 + \sqrt{(5.716 - 6.170)^2 + 4(.977)^2 6.170(5.716)}}{2(2.484)(2.391).977}$$

$$A = \frac{11.919}{12.503}$$

$$A = .962.$$

Again 16 items were identified as falling beyond the .75 cutoff. However, since these items were not identical to the 16 removed in the first cycle a third iteration of the algorithm was performed which produced a line with the slope

$$A = \frac{6.100 - 6.698 + \sqrt{(6.100 - 6.698)^2 + 4(.978)^2 6.698(6.100)}}{2(2.588)(2.470).978}$$

$$A = \frac{11.919}{12.503}$$

$$A = .953.$$

This calculation resulted in the removal of 15 items. The remaining 85 items were used as the basis for the fourth and final iteration

producing a line with the slope

$$A = \frac{6.059 - 6.709 + \sqrt{(6.059 - 6.709)^2 + 4(.977)^2 6.709(6.059)}}{2(2.590)(2.461).977}$$

$$A = \frac{11.825}{12.455}$$

$$A = .949.$$

The 15 items identified as potentially biased in the Group II, Group III comparison along with their respective distance from the line of best fit in delta units are provided in Table 6. Positive values indicate a bias in favor of Group II whereas negative values indicate a bias in favor of Group III. Both groups are composed of candidates other than blacks participating in the July 1981 biochemistry-physiology National Board dental examination.

Items Identified as Biased by the Chi Square Procedure

Chi square values for the 100 test items were obtained for the three comparisons following Scheuneman's modified chi square technique. Table 7 lists each item along with its corresponding chi square value for each of the three comparisons. Items with a chi square value reaching the .05 level of confidence ($\geq$ 5.99) are indicated with an asterisk. At this level of confidence the Group I, Group II comparison identified six potentially biased items, the Group I, Group III comparison identified ten potentially biased items, and the Group II, Group III found no biased items. As an alternate measure for identifying potentially biased items, rank ordered chi square values with their respective item numbers are provided in Table 8.

Table 6

Items Identified as Biased in the Group II, Group III Delta Plot Comparison With a .75 Cutoff

Item Number	Distance From the Line
11	-1.17
23	-1.37
24	-0.89
30	1.00
32	-1.12
33	-1.07
37	0.93
42	1.34
46	0.87
57	0.82
61	-1.13
62	0.80
92	-0.96
95	2.42
98	0.77

Table 7

Chi Square Values by Item for Each of the Three Comparisons

Item Number	Chi Square Value Group I, Group II Comparison	Group I, Group III Comparison	Group II, Group III Comparison
1	0.07	0.26	0.07
2	0.17	0.05	0.23
3	0.43	0.36	0.21
4	1.20	1.81	0.24
5	0.16	0.67	0.58
6	0.00	0.03	0.03
7	0.20	0.11	0.46
8	0.50	0.68	0.21
9	2.54	1.81	0.10
10	0.29	0.11	0.23
11	0.27	0.33	0.32
12	0.10	0.46	0.34
13	0.67	1.84	0.39
14	0.33	0.45	0.55
15	1.89	1.72	0.32
16	0.60	2.19	1.12
17	0.49	0.57	0.18
18	0.30	0.18	0.34
19	1.15	1.98	0.16
20	0.86	0.19	0.53
21	0.74	0.28	0.37
22	0.28	1.04	0.31
23	0.20	0.20	0.46
24	1.23	3.25	3.79
25	1.83	0.53	0.62
26	1.25	1.82	0.17
27	0.07	0.17	0.12
28	1.83	1.33	0.16
29	3.51	4.29	0.43
30	1.06	0.23	0.33
31	1.94	0.39	2.12
32	0.39	0.11	0.36
33	3.86	10.87*	4.07
34	0.87	0.54	0.09
35	7.74*	6.34*	0.87
36	3.81	1.22	0.75
37	1.29	6.58*	2.36
38	0.72	0.77	0.01
39	0.47	1.93	2.74
40	0.14	1.19	1.51
41	1.48	1.57	0.88

Table 7 (continued)

Item Number	Chi Square Value Group I, Group II Comparison	Group I, Group III Comparison	Group II, Group III Comparison
42	0.49	3.79	2.82
43	0.76	1.48	0.42
44	3.51	8.20	2.61
45	1.47	1.22	0.88
46	0.98	4.27	1.66
47	8.45*	2.98	2.88
48	0.97	0.29	0.27
49	5.14	4.36	0.70
50	1.13	1.33	3.43
51	1.65	2.97	2.58
52	1.20	0.70	0.85
53	0.90	3.55	0.98
54	5.40	3.36	0.59
55	0.24	1.77	0.75
56	0.63	0.17	0.19
57	0.30	1.88	1.22
58	2.32	3.08	0.69
59	1.11	0.75	1.73
60	0.05	0.38	0.17
61	1.96	6.55*	2.24
62	1.90	0.23	1.38
63	2.97	3.80	2.64
64	6.31*	4.75	3.22
65	5.73	3.23	1.81
66	1.49	1.30	0.23
67	0.21	0.28	0.13
68	0.46	1.26	1.26
69	3.71	2.46	0.43
70	5.87	2.57	1.53
71	3.09	8.93*	2.02
72	0.70	0.58	0.14
73	0.14	2.14	1.35
74	3.99	3.50	0.52
75	3.99	1.37	1.17
76	0.30	0.48	0.21
77	0.25	0.81	0.62
78	1.05	0.14	1.23
79	1.87	1.53	0.59
80	2.75	1.53	1.76
81	1.21	1.13	1.58
82	3.78	5.97	1.65
83	0.29	0.16	0.20
84	2.16	1.47	0.49
85	8.95*	8.27*	1.41

Table 7 (continued)

Item Number	Chi Square Value Group I, Group II Comparison	Group I, Group III Comparison	Group II, Group III Comparison
86	1.53	1.17	1.69
87	5.07	7.76*	1.83
88	2.12	1.99	2.50
89	0.18	0.54	0.21
90	0.22	0.35	1.05
91	7.29*	7.36*	1.89
92	0.17	0.24	0.06
93	0.47	0.53	0.17
94	0.89	0.53	0.23
95	0.22	0.33	0.65
96	0.21	0.38	0.23
97	1.42	1.45	0.67
98	0.41	2.59	1.19
99	13.45*	9.59*	0.47
100	0.45	0.22	0.50

*Significant at the .05 level of confidence with 2 degrees of freedom.

Table 8

Rank Ordered Chi Square Values for Each of the Three Comparisons

	Comparison					
	Group I, Group II		Group I, Group III		Group II, Group III	
Rank	X^2	Item	X^2	Item	X^2	Item
1	13.45	99	10.97	33	4.07	33
2	8.95	85	9.59	99	3.79	24
3	8.45	47	8.93	71	3.43	50
4	7.74	35	8.27	85	3.22	64
5	7.29	91	8.20	44	2.88	47
6	6.31	64	7.76	87	2.82	42
7	5.87	70	7.36	91	2.74	39
8	5.73	65	6.58	37	2.64	63
9	5.40	54	6.55	61	2.61	44
10	5.14	49	6.34	35	2.58	51
11	5.07	87	5.97	82	2.50	88
12	3.99	75	4.75	64	2.36	37
13	3.99	74	4.36	49	2.24	61
14	3.86	33	4.29	29	2.12	31
15	3.81	36	4.27	46	2.02	71
16	3.78	82	3.80	63	1.89	91
17	3.71	69	3.79	42	1.83	87
18	3.51	29	3.55	53	1.18	65
19	3.51	44	3.50	74	1.76	80
20	3.09	71	3.36	54	1.73	59
21	2.97	63	3.25	24	1.69	86
22	2.75	80	3.23	65	1.66	46
23	2.54	9	3.08	58	1.65	82
24	2.32	58	2.98	47	1.58	81
25	2.16	84	2.97	51	1.53	70
26	2.12	88	2.59	98	1.41	85
27	1.96	61	2.57	70	1.38	62
28	1.94	31	2.46	69	1.35	73
29	1.90	62	2.19	16	1.26	68
30	1.89	15	2.14	73	1.23	78
31	1.87	79	1.99	88	1.22	57
32	1.83	25	1.98	19	1.19	98
33	1.83	28	1.93	39	1.17	75
34	1.65	51	1.88	57	1.15	40
35	1.53	86	1.84	13	1.12	16
36	1.49	66	1.82	26	1.05	90
37	1.48	41	1.81	9	0.98	53
38	1.47	45	1.81	4	0.88	41
39	1.42	97	1.77	55	0.88	45
40	1.29	37	1.72	15	0.87	35
41	1.25	26	1.57	41	0.85	52

Table 8 (continued)

	Comparison					
	Group I, Group II		Group I, Group III		Group II, Group III	
Rank	X^2	Item	X^2	Item	X^2	Item
42	1.23	24	1.53	79	0.75	36
43	1.22	81	1.53	80	0.75	55
44	1.20	4	1.48	43	0.70	49
45	1.20	52	1.47	84	0.69	58
46	1.15	19	1.45	97	0.67	97
47	1.13	50	1.37	75	0.65	95
48	1.11	59	1.33	50	0.62	77
49	1.06	30	1.33	28	0.62	25
50	1.05	78	1.30	66	0.59	79
51	0.98	46	1.26	68	0.59	54
52	0.97	48	1.22	45	0.58	5
53	0.90	53	1.22	36	9,55	14
54	0.87	34	1.19	40	0.53	20
55	0.87	94	1.17	86	0.52	74
56	0.86	20	1.13	81	0.50	100
57	0.76	43	1.04	22	0.49	84
58	0.74	21	0.81	77	0.47	99
59	0.72	38	0.77	38	0.46	7
60	0.70	72	0.75	59	0.46	23
61	0.67	12	0.70	52	0.43	29
62	0.63	56	0.68	8	0.43	69
63	0.60	16	0.67	5	0.42	43
64	0.50	8	0.58	72	0.39	13
65	0.49	17	0.57	17	0.37	21
66	0.49	42	0.54	34	0.36	32
67	0.47	39	0.54	89	0.34	12
68	0.47	93	0.53	25	0.34	18
69	0.46	68	0.53	94	0.33	30
70	0.45	100	0.53	93	0.32	11
71	0.43	3	0.48	76	0.32	15
72	0.41	98	0.46	12	0.31	22
73	0.39	32	0.45	14	0.27	48
74	0.33	14	0.39	31	0.24	4
75	0.30	18	0.38	60	0.23	10
76	0.30	76	0.38	96	0.23	66
77	0.30	57	0.36	3	0.23	96
78	0.29	83	0.35	90	0.23	94
79	0.29	10	0.33	95	0.23	2
80	0.28	22	0.33	11	0.21	76
81	0.27	11	0.29	48	0.21	3
82	0.25	77	0.28	21	0.21	8
83	0.24	55	0.28	67	0.21	89
84	0.22	95	0.26	1	0.20	83

Table 8 (continued)

	Comparison					
	Group I, Group II		Group I, Group III		Group II, Group III	
Rank	X^2	Item	X^2	Item	X^2	Item
85	0.22	90	0.24	92	0.19	56
86	0.21	96	0.23	62	0.18	17
87	0.21	67	0.23	30	0.17	60
88	0.20	23	0.22	100	0.17	26
89	0.20	7	0.20	23	0.17	93
90	0.18	89	0.19	20	0.16	28
91	0.17	92	0.18	18	0.16	19
92	0.17	2	0.17	56	0.14	72
93	0.16	5	0.17	27	0.13	67
94	0.14	73	0.16	83	0.12	27
95	0.14	40	0.14	78	0.10	9
96	0.10	12	0.11	10	0.09	34
97	0.07	1	0.11	7	0.07	1
98	0.07	27	0.11	32	0.06	92
99	0.05	60	0.05	2	0.03	6
100	0.00	6	0.03	6	0.01	38

Items Identified as Biased by Both the Delta Plot and the Chi Square Procedure

Items found biased by both the delta plot procedure and the chi square procedure for each of the three comparisons are given in Table 9.

Table 9

Items Found Biased by Both the Delta Plot Procedure and the Chi Square Procedure for Each of the Three Comparisons

Comparison		
Group I, Group II	Group I, Group III	Group II, Group III
# 35	# 33	none
# 47	# 35	
# 85	# 37	
# 91	# 61	
# 99	# 71	
	# 85	
	# 87	
	# 91	
	# 99	

CHAPTER V

DISCUSSION

Items Identified as Biased by the Delta Plot Procedure

After reviewing the results of Sinnott's modification of the delta plot procedure there is question as to whether a distance of .75 delta units from the line of best fit was an appropriate cutoff to employ with these data. Given the large number of items designated as potentially biased with the .75 cutoff it appears to be unlikely. Sinnott's (1980) discussion of the necessity of an empirical determination of the cutoff used for each set of data is exemplified in this study. A cutoff of .75 resulted in the elimination of too many items during the calculation of lines of best fit, i.e., 38 items in the Group I, Group II comparison, 45 items in the Group I, Group III comparison, and 15 items in the Group II, Group III comparison. Consequently the .75 cutoff was not particularly useful in identifying items that significantly departed from the line of best fit in relation to the other test items. In the case of these data it appears that selection of a larger cutoff for empirical evaluation is warranted. Increasing the cutoff to 1.50 delta units from the line reduces the number of items identified as potentially biased considerably. Those items identified as potentially biased using a 1.50 cutoff are listed in Tables 10-12.

Inspection of Tables 10 through 12 reveals 12 items found to be potentially biased in the Group I, Group II comparison, 11 items

Table 10

Items Identified as Biased in the Group I, Group II Delta Plot Comparison With a 1.50 Cutoff

Item Number	Distance From the Line
9	-1.63
30	-1.98
32	1.62
35	2.25
36	1.88
54	-1.59
75	1.87
85	1.89
91	-1.55
92	1.94
95	-1.73
99	-1.84

Table 11

Items Identified as Biased in the Group I, Group III Delta Plot Comparison With a 1.50 Cutoff

Item Number	Distance From the Line
1	1.72
26	-1.57
33	-2.44
35	2.10
37	1.72
61	-2.07
85	1.79
87	1.59
91	-1.65
92	-2.62
99	-1.50

Table 12

Items Identified as Biased in the Group II, Group III Delta Plot Comparison With a 1.50 Cutoff

Item Number	Distance From the Line
95	2.42

found to be potentially biased in the Group I, Group III comparison, and one item found to be potentially biased in the Group II, Group III comparison. These results are most plausible than those obtained with the .75 cutoff.

The results of initial concern are those of the homogeneous group comparison. The comparison between Groups II and III was performed to establish a baseline for these data and technically should yield no difference in the performance between the groups. The single item identified as potentially biased in this comparison is an easy item with p values of .95, .98 and .89 respectively for the three groups. This would disqualify the item as biased, therefore yielding no potentially biased items in the Group II, Group III baseline comparison.

In spite of the use of a 1.50 cutoff point both the Group I, Group II and the Group I, Group III comparisons still identified 12 and 11 items respectively as being potentially biased. However, the design of this study provides an additional source of information concerning the black student/other student comparison by including two such comparisons from which to extrapolate. Examining the data from both comparisons identifies only five items as potentially biased in both instances. These items are numbers 35, 85, 91, 92 and 99. Among these five items, item 92 may be disqualified as biased because it has p values of .94, .98 and .99 for the three groups and is therefore too easy an item.

Taking into account the larger cutoff distance of 1.50 delta units from the line of best fit, and a selection of only those items

identified as biased in both the Group I, Group II and the Group I, Group III comparisons, the delta plot procedure identified items 35, 85, 91 and 99 of the July 1981 biochemistry-physiology National Board dental examination as being potentially biased. Before attempting to determine possible reasons these items were identified as biased the results of the chi square procedure will be discussed.

Items Identified as Biased by the Chi Square Procedure

The results obtained using the chi square procedure are more straightforward and less subject to interpretation than the results obtained using the delta plot procedure. As Shepard, et.al. (1980) noted, the chi square method is not based on arbitrary rules. Once the level of significance is selected the chi square values are calculated and referenced in the chi square table under the appropriate degrees of freedom. A significance level of .05 is generally accepted in the social sciences and does not require justification. In this study a .05 level of significance identified items 35, 47, 64, 85, 91 and 99 as potentially biased in the Group I, Group II comparison and items 33, 37, 44, 61, 71, 85, 87, 91 and 99 as potentially biased in the Group I, Group III comparison. There were no chi square values reaching the .05 level of significance in the Group II, Group III baseline comparison. If one wished, a .01 level of significance (9.21 with two degrees of freedom) could be employed to increase certainty of item selection. Using a .01 level of significance with these data identifies one item, number 99, as biased in the Group I, Group II comparison and two items, numbers 33 and 99, as biased in the

Group I, Group III comparison. In this study data obtained using the .05 level of significance will be used.

Since the groups used in this study are approximately equal and cell frequencies are relatively small, rank ordering chi square values, as suggested by Schueneman (1979), does not provide additional insight into the identification of biased items with these data. It is with unequal groups and large cell frequencies that the modified chi square values are least likely to approximate the chi square distribution. Examination of Table 8 shows no difference in the identification of biased items from those identified using a probability measure. Once the chi square values have been rank ordered the selection of items for inspection is the subjective decision of the investigator. In this study rank ordering of chi square values merely serves as a check for the data derived with the test of significance.

As with the data from the delta plot, the nature of the design of this study provides two measures of black student/other student comparisons using the chi square procedure. Examination of data from both comparisons identifies four items as potentially biased in both cases. These items are numbers 35, 85, 91 and 99.

Items Identified as Biased by Both the Delta Plot and the Chi Square Procedures

From the onset, the purpose of this study has been to determine the extent to which the delta plot and the chi square methods for detecting test item bias agree in their detection of biased items on the biochemistry-physiology examination of the July 1981 National Board dental examinations. Reference to Table 9 shows four items which were identified as biased by both methods. The four

items are numbers 35, 85, 91 and 99. It is noteworthy that these four items are the same four items identified as biased for both black student/other student comparisons within each method. Items 35 and 85 were identified as biased in favor of the black student group. Items 91 and 99 were identified as biased in favor of the other student group. Since the four items identified as biased by both methods when using the reliability checks made available within each of the methods are identical, there is evidence the level of agreement between the two methods is substantial. The final task of this study is to attempt a determination of why these four items were identified as biased by the two methods.

Possible Reasons the Items Were Identified as Biased by the Two Methods

Lois Burrill (1982) states there are many reasons an item may be identified as biased using any method of item bias detection. The reasons may include the placement of the item on the page, the order of distractors, or any other aspect of test format. Indeed, on occasion these have been found to have more to do with item bias detection than the content of the item itself (Schueneman, 1978). The point being made is the necessity of examining all possible, if not logical, reasons an item may have been identified as biased. What follows is an examination of the reasons items 35, 85, 91 and 99 of the National Board biochemistry-physiology examination were identified as biased by both the delta plot and the chi square procedures. In addition to the four items identified as biased by both methods, items 47 and 33 will be included in the analysis. These items were

chosen because item 47 was identified as biased by both methods but only in the Group I, Group II comparison and item 33 was identified as biased by both methods but only in the Group I, Group III comparison. Item 47 was identified as biased in favor of the black student group. Item 33 was identified as biased in favor of the other student group. It is hoped that inclusion of these items in the analysis will reveal if any difference exists between the two black student/other student comparisons. It should be noted that selection of these items was based on the original results of the study using .75 as a distance cutoff for the delta plot procedure.

Content: When an item is identified as biased the first logical assumption is that there is something in the content of the item which results in one group performing significantly better than another group, provided differences in ability have been accounted for in advance. With this in mind the content of each of the six items under analysis was reviewed by a dental expert on staff at the American Dental Association. The expert first organized the items according to topic and found items 91 and 35 related to biochemistry while items 33, 47, and 99 were related to physiology. Item 85 could have been classified as either topic. Keeping in mind that items 33, 91 and 99 were identified as biased in favor of the other student group, and items 35, 47 and 85 were identified as biased in favor of the black student group, there does not appear to be any difference in the performance of the two groups by topic.

A more specific examination of content of these items found items 47, 85 and 91 to be distantly related because they all pertained

to some aspect of the urinary tract. Other than this distant relationship the dental expert found nothing in the wording of these three items that would be sensitive to either of the groups. Based on this he concluded these three items were not biased based on content. He noted item 35 could conceivably be biased in favor of the black student group because of the mention of the dark brown or black pigment of the skin, melanin. Black students may be more attuned to this subject area.

The dental expert found items 33 and 99 to be closely related in content area. He stated the content in these two items was so closely related it was likely to have been included in the same lecture. Although he was unable to find anything in the wording that would be sensitive to either group, he did believe content could be a source of bias in favor of the other student group for these two items, given their close relationship.

In summary, the dental expert believed items 47, 85 and 95 were not biased based on content. Item 35 could possibly be biased in favor of blacks based on content. Items 33 and 99 were likely to be identified as biased in favor of the other student group based on content.

Sample: The question which logically follows is whether the sample used in the study was in any way responsible for the identification of the items as biased. It has been noted that dental schools occasionally cover different subject areas in varying degrees. Because items 33 and 99 cover the same topic it is possible this topic was covered more extensively at certain schools. If content coverage

within a school was responsible for a difference in item performance it would be a result of the sample employed.

The sample selection for the black student group in this study was not random due to the limited black student pool. Consequently 63 percent of the students in the black student group came from the two schools which have the highest black enrollment of the 60 dental schools. Of the 63 percent, 25 percent came from School I and 38 percent came from School II. With such a high concentration of students coming from two schools it is necessary to determine whether the students from these two schools performed significant differently on the six items from the remaining 37 percent of black students coming from other dental schools. Table 13 provides the percent of students answering each item correctly from School I, School II and all Other Schools. Also included in the table is the total number of students in Group I answering the items correctly.

These data show School I students performing consistently lower than School II and all Other School students on all items except item 85 where all schools performed approximately equal. School II students performed well in comparison to the Other Schools group and would therefore not account for a lower proportion of Group I students answering the items correctly. Based on the data presented in Table 13 it does not appear that either School I or School II students disproportionately affect the total group score.

Item Difficulty and Discrimination: Items which are too difficult or have a poor discriminating ability affect the way in which the item is answered. These items tend to elicit more randomness of choice

Table 13

Performance of Group I by School

Item	Percent Answering Correctly			
	School I	School II	Other Schools	All Schools
33	9%	17%	19%	45%
35	20%	35%	26%	81%
47	14%	26%	20%	60%
85	19%	16%	23%	58%
91	5%	17%	10%	32%
99	2%	10%	8%	20%

than more statistically sound items and could lead to a misrepresentation of item bias. Table 14 shows the national average for each of the six items as well as the percent answering the item correctly for both the high and low performing groups on this examination. All items had a good discriminating power and were not particularly difficult for the national group participating in this examination. For the six items under analysis in this investigation neither item difficulty or item discrimination ability appear to be a functioning variable in their identification as biased.

Probability: The probability that a certain number of items would be identified as biased on any test must be considered in this analysis. However, given the results indicated by the data under analysis it is unlikely that probability was responsibile for the selection of the biased items in this study. The fact that the four items were identified as biased by both methods for both black student/other student comparisons argues against their being selected by chance. Rather, this provides a measure of verification that the items were identified as biased for some other reason. Since items 33 and 47 were identified as biased by both methods but for only one black student/other student comparison it is more likely that probability was responsible for their selection. However, again, identification by both methods makes this assumption questionable. When in doubt additional information such as the content analysis provided by the dental expert should be employed in any decision making.

Format: The format of the examination as well as each item was reviewed by one of the editors of the National Board dental examinations.

Table 14

National Statistics for the
Six Items Examined for Bias

Item	National Average	Percent of High Group Answering Correctly	Percent of Low Group Answering Correctly
33	65.5%	87%	42%
35	65%	85%	45%
47	47.5%	67%	28%
85	47.5%	69%	26%
91	49.5%	74%	25%
99	48%	71%	25%

She was unable to find any aspect of the format which would have accounted for selection of the items as biased.

Conclusion

The results of this study support the hypothesis that the delta plot procedure and the chi square procedure would detect a similar number and a similar set of biased items on the same examination. Indeed, the two methods detected an identical number and set of items as biased when both black student/other student comparisons were used as an additional control. The inclusion of items 47 and 33 in the final analysis did not yield any information on a difference in performance between the Group I, Group II and the Group I, Group III comparisons. The inclusion of these items was beneficial however, because item 33 was thought to be content biased by the dental expert.

Future investigations using data from National Board dental examinations may be helpful considering the results of this study. Particularly, a re-examination of items 33, 35 and 99 on a future examination seems advisable to confirm or disconfirm any difference in the performance between the black student group and the other student group. This would substantiate whether the delta plot and chi square procedures were appropriate methods to use with National Board data.

Items 47, 85 and 91 might also be included in a future examination with changes in format such as placement in the examination or arrangement of distractors to determine whether these artifacts were

responsible for their identification as biased. Format changes for items 33, 35 and 99 would be an advisable added precaution.

Another reason a future study using data from the National Board dental examination is warranted stems from the spurious results obtained in this study using the inappropriately low .75 distance cutoff in the delta plot procedure. An investigation empirically validating a distance measure is necessary to confirm the results of the present study as well as to provide additional information regarding test item bias in National Board dental examinations.

According to the results of this study reduction in the number of items included in the final analysis for bias can be accmplished by using two item bias detection procedures and two group comparisons within each method. In this study the number of biased items was reduced to four when all comparative conditions were taken into account. The additional control lends some reassurance that the items identified as biased have some variable at work which effects the performance of a particular group. However, reduction of the number of items for review may omit a potentially biased item from the analyais. In this study if item 33 had been omitted from the analysis the content would not have been scrutinized for bias. Where time allows it appears that all items identified by both methods should be examined for sources of bias.

CHAPTER VI

SUMMARY

A study was conducted to investigate the amount of agreement between two methods for detecting test item bias. Data from the July 1981 biochemistry-physiology National Board dental examination was used to test the hypothesis that the delta plot procedure and the chi square procedure for detecting test item bias would identify a similar number and set of biased items. Results indicated that the two methods of item bias detection have a high level of agreement in the identification of biased items.

The items identified as biased in the study were examined for possible source of bias. It was decided that for three of the six items identified, bias was produced by a factor not identifiable by the methods used in this study. For the three remaining items it was thought that some aspect of content could be responsible for item bias. An additional study of these items is necessary before a definite decision can be made regarding a source of bias.

These results support the use of the two item bias detection procedures with natural test data and a relatively small sample size. The results also suggest a similar set of items will be identified as biased by the two methods provided a suitable distance cutoff is employed with the delta plot procedure.

Bibliography

Angoff, W.H. The investigation of test bias in the absence of an outside criterion. Paper presented at the National Institute of Education Conference on Test Bias, Maryland, December 1975.

Angoff, W.H. The use of difficulty and discrimination indices in the identification of biased test items. Paper presented at the Johns Hopkins University Symposium on Education Research, "Test Item Bias Methodology: State of the Art," Washington, D.C., November 1980.

Angoff, W.H., & Ford, S.F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-105.

Angoff, W.H., & Herring, C.L. A study of the appropriateness of the Law School Admission Test for Canadian and American students. Princeton, N.J.: Educational Testing Service, 1971.

Angoff, W.H., & Sharon, A.T. The evaluation of differences in test performance of two or more groups. Educational and Psychological Measurement, 1974, 34, 807-816.

Burrill, L.E. Comparative studies of item bias methods. In Ronald A. Berk (ed.), Handbook of methods for dictating test bias. Baltimore: The Johns Hopkins University Press, 1982.

Cardall, C., & Coffman, W.E. A method for comparing the performance of different groups on the items in a test. College Board Research and Development Reports 64-5, No. 9 and Educational Testing Service Research Bulletin 64-63. Princeton, N.J.: Educational Testing Service, 1964.

Carlton, S.T., & Marco, G.L. Methods used by Educational Testing Service Testing Programs for detecting and eliminating item bias. Paper presented at the Johns Hopkins University Symposium on Educational Research, "Test Item Bias Methodology: State of the Art," Washington, D.C., November 1980.

Cleary, T.A., & Hilton, T.L. An investigation of item bias. Educational and Psychological Measurement, 1968, 28, 71-75.

Cole, N.S. Approaches to examining bias in achievement test items. Paper presented at the national meeting of the American Personnel and Guidance Association, Washington, D.C., March 1978.

Cole, N.S. Scientific standards for tests and social values. Paper presented at the annual meeting of the American Psychological Association, Montreal, September 1980.

Cole, N.S. Bias in testing. American Psychologist, 1981, 36, 1067-1077.

Hunter, J.E. A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items. Paper presented at the National Institute of Education Conference on Test Bias, Maryland, December 1975.

Information sheet for the National Board dental examinations.

Ironson, G.H., & Subkoviak, M.J. A comparison of several methods of assessing item bias. Journal of Educational Measurement, 1979, 16, 209-225.

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

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

Merz, W.R., & Grossen, N.E. An empirical investigation of six methods for examining test item bias. Report submitted to the National Institute of Education, Grant NIE-6-78-0067, California State University, Sacramento, California, 1979.

Rudner, L.M. Individual assessment accuracy. Paper presented at the annual meetings of the American Research Association and the National Council on Measurement in Education, New York, March 1982.

Rudner, L.M. An evaluation of select approaches for biased item identification. Unpublished doctoral dissertation, Catholic University of America, 1977b.

Rudner, L.M., & Getson, P.R. Item bias research has its limitations. Paper presented at the annual meeting of the National Council for Measurement in Education, New York, March 1982.

Rudner, L.M., Getson, P.R., & Knight, D.L. Biased item detection techniques. Journal of Educational Statistics, 1980, 5, 213-233.

Rudner, L.M , Getson, P.R., & Knight, D.L. A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 1980, 17, 1-10.

Scheuneman, J. A method of assessing bias in test items. Journal of Educational Measurement, 1979, 16, 143-152.

Scheuneman, J. Consistency across administrations of certain indices of bias in test items. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Scheuneman, J.D. Item bias and test scores. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, March 1982.

Shepard, L.A. Definitions of bias. Paper presented at the Johns Hopkins University Symposium on Educational Research, "Test Item Bias Methodology: State of the Art," Washington, D.C., November 1980.

Shepard, L.A., Camilli, G., & Averill, M. Comparison of six procedures for detecting test item bias using both internal and external ability criteria. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, April 1980.

Sinnott, L.A. Differences in item performance across groups. Report for the Graduate Management Admission Council 80-1, Educational Testing Service Research Report RR-80-19. Princeton, N.J.: Educational Testing Service, 1980.

Stricker, L.J. A new index of differential subgroup performance: Application to the GRE Aptitude Test. GRE Board Professional Report GREB No. 78-7P, Educational Testing Service Research Report 81-13. Princeton, N.J.: Educational Testing Service, 1981.

Subkoviak, M.J., Mack, J.S., & Ironson, G.H. Item bias detection procedures: Empirical validation. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981.

APPENDIX A

FORTRAN PROGRAM USED TO COMPUTE DELTA VALUES AND THEIR PLOTS

```
INFILE IN1:
INPUT GROUP 77 #4;
   DATA RAW13;
INFILE IN2;
INPUT (X1-X50) (1.0) #2 (X51-X100) (1.0);
   DATA ALL1;
MERGE ELAINE2 RAW13;
   PROC SORT DATA-ALL;
BY GROUP:
   PROC MEANS NOPRINT DATA=ALL1;
VAR X1-X100;
BY GROUP;
OUTPUT OUT=ALL2 MEAN=P1-P100;
   DATA ALL 3;
SET ALL2;
DROP GROUP:
   PROC MATRIX;
FETCH X DATA=ALL3;
TR=X';
OUTPUT TR OUT=TEMP1;
   DATA SAVE.D13;
SET TEMP1;
G1=4*PROBIT(COL1)+13;
G2=4*PROBIT(COL2)+13;
G3=4*PROBIT(COL3)+13;
DROP COL1-COL3;
END OF DATA
READY
```

APPENDIX B

FORTRAN PROGRAM USED TO COMPUTE DISTANCE FOR LINE OF BEST FIT

```
    DATA ALL1
SET SAVE.D13
IF G2=. OR G1=. THEN DELETE;
DROP G3
    DATA ALL2
SET ALL1;
Y=.981*G1+.193;
D=(Y-G2)/1.401;
DROP G1;
    DATA ALL3;
SET ALL2;
IF -.75<=D AND D<=.75 THEN DELETE;
    PROC PRINT DATA=ALL3;
END OF DATA
READY
```

APPENDIX C

FORTRAN PROGRAM USED TO COMPUTE SLOPE FOR LINE OF BEST FIT

```
    DATA ALL1;
SET SAVE.D13;
IF G2=. OR G1=. THEN DELETE
DROP G3;
    PROC CORR DATA=ALL1 NOSIMPLE;
VAR G1 G2;
    PROC MEANS DATA=ALL1 N MEAN STD VAR MAXDEC=3;
VAR G1 G2;
END OF DATA
READY
```

APPENDIX D

FORTRAN PROGRAM USED TO DELETE ITEMS FOR LINE OF BEST FIT

```
    DATA ALL1;
SET SAVE.D13;
IF G2=. or G1=. THEN DELETE;
IF ROW='ROW2' THEN DELETE:
IF ROW='ROW5' THEN DELETE;    ETC.
DROP G3;
    PROC CORR DATA=ALL1 NO SIMPLE;
VAR G1 G2;
    PROC MEANS DATA=ALL1 N MEAN STD VAR MAXDEC=3;
VAR G1 G2;
END OF DATA
READY
```

APPENDIX E

FORTRAN PROGRAM USED TO COMPUTE CHI SQUARE VALUES

```
DIMENSION X(350,100),GROUP(350),SCORE(350),TOTGP(3)
DIMENSION TOTINT(3),INTERV(350,3),INTGP(3,3)XINTGP(100,3,3)
DIMENSION E(100,3,3),RINTGP(3,3),CHI(100,3),FLAG(100,3)
REAL RINTGP,CHI,E
INTEGER X,GROUP,SCORE,TOTGP,TOTINT,INTERV,INTGP,XINTGP,FLAG
READ(10,1) ((X(I,J),J=1,100),I=1,350)
READ(11,2) (GROUP(I),I=1,350)
DO 10 I=1,350
SCORE(I)=0
DO 10 J=1,100
SCORE(I)=SCORE(I)+X(I,J)
CONTINUE
DO 15 I=1,3
TOTGP(I)=0
TOTINT(I)=0
DO 20 I=1,350
DO 20 J=1,3
INTERV(I,J)=0
DO 25 I=1,350
IF (SCORE(I) .GE. 46 .AND. SCORE(I) .LE. 59) GO TO 21
IF (SCORE(I) .GE. 60 .AND. SCORE(I) .LE. 68) GO TO 22
IF (SCORE(I) .GE. 69 .AND. SCORE(I) .LE. 83) GO TO 23
INTERV (I,3)=1
TOTINT(3)=TOTINT(3)+1
GO TO 25
 INTERV(I,2)=1
TOTINT(2)=TOTINT(2)+1
GO TO 25
 INTERV(I,1)=1
TOTINT(1)=TOTINT(1)+1
CONTINUE
DO 30 I=1,350
IF (GROUP(I) .EQ. 1) GO TO 27
IF (GROUP(I) .EQ. 2) GO TO 28
IF (GROUP(I) .EQ. 3) GO TO 29
TOTGP(1)=TOTGP(1)+1
GO TO 30
TOTGP(2)=TOTGP(2)+1
GO TO 30
TOTGP(3)=TOTGP(3)+1
CONTINUE
DO 35 I=1,3
DO 35 J=1,3
INTGP(I,J)=0
DO 40 J=1,3
DO 40 I=1,350
IF (INTERV(I,J) .EQ. 1 .AND. GROUP(I) .EQ. 1) GO TO 36
IF (INTERV(I,J) .EQ. 1 .AND. GROUP(I) .EQ. 2) GO TO 37
```

```
IF (INTERV(I,J) .EQ. 1 .AND. GROUP(I) .EQ. 3) GO TO 38
GO TO 40
INTGP(J,1)=INTGP(J,1)+1
GO TO 40
INTGP(J,2)=INTGP(J,2)+1
GO TO 40
INTGP(J,3)=INTGP(J,3)+1
CONTINUE
DO 45 I=1,100
DO 45 J=1,3
DO 45 K=1,3
XINTGP(I,J,K)=0
DO 60 I=1,100
DO 60 K=1,350
IF (X(K,I) .NE. 1) GO TO 60
DO 60 J=1,3
IF (INTERV(K,J) .EQ. 1 .AND. GROUP(K) .EQ. 1) GO TO 51
IF (INTERV(K,J) .EQ. 1 .AND. GROUP(K) .EQ. 2) GO TO 52
IF (INTERV(K,J) .EQ. 1 .AND. GROUP(D) .EQ. 3) GO TO 53
GO TO 60
XINTGP(I,J,1)=XINTGP(I,J,1)+1
GO TO 60
XINTGP(I,J,2)=XINTGP(I,J,2)+1
GO TO 60
XINTGP(I,J,3)=XINTGP(I,J,3)+1
CONTINUE
DO 65 I=1,100
DO 65 J=1,3
DO 65 K=1,3
E(I,J,K)=0.
DO 90 I=1,100
DO 90 L=1,3
GO TO (66,67,68),L
K=1
KK=2
GO TO 70
K=1
KK=3
GO TO 70
K=2
KK=3
DO 75 J=1,3
INTOT=INTGP(J,K)+INTGP(J,KK)
RINTGP(J,K)=INTGP(J,K)
RINTGP(J,KK)=INTGP(J,KK)
E(I,J,K)=(RINTGP(J,K)*(XINTGP(I,J,K)+XINTGP(I,J,KK)))/INTOT
E(I,J,KK)=(RINTGP(J,KK)*(XINTGP(I,J,K)+XINTGP(I,J,KK)))/INTOT
CONTINUE
CHI(I,L)=0.
DO 80 J=1,3
E1=(E(I,J,K)-XINTGP(I,J,K))**2/E(I,J,K)
```

```
E2=(E(I,J,KK)-XINTGP(I,J,KK))**2/E(I,J,KK)
CHI(I,L)=CHI(I,L)+E1+E2
CONTINUE
WRITE(6,4)I,J,K,KK,(E,I,J,K),E(I,J,KK),J=1,3)
FORMAT(2X,414,6F10.2)
CONTINUE
DO 95 I=1,100
DO 95 J=1,3
IF (CHI(I,J) .GT. 5.99) GO TO 93
FLAG(I,J)=0
GO TO 95
FLAG(I,J)=1
CONTINUE
DO 100 I=1,100
WRITE(6,3) I,(CHI(I,J),FLAG(I,J),J=1,3)
CONTINUE
FORMAT(50I1,/,50I1)
FORMAT(76X,I1,///)
FORMAT(2X,I3,3(2X,F8.3,3X,I1))
STOP
END
```

APPROVAL SHEET

The thesis submitted by Elaine Kopera has been read and approved by the following committee:

Dr. Jack A. Kavanagh, Director
Associate Dean and Associate Professor, School of Education, Loyola

Dr. Anne M. Juhasz
Professor, Foundations of Education, Loyola

The final copies have been examined by the director of the thesis and the signature which appears below verifies the fact that any necessary changes have been incorporated and that the thesis is now given final approval by the Committee with reference to content and form.

The thesis is therefore accepted in partial fulfillment of the requirements for the degree of Master of Arts.

Dec. 6, 1982
Date

Jack A. Kavanagh
Director's Signature