1983

Investigation of Several Procedural Modifications to Delta Plot Methodology

Rita Karwacki Bode
Loyola University Chicago

INVESTIGATION OF SEVERAL PROCEDURAL MODIFICATIONS TO DELTA PLOT METHODOLOGY

by

Rita Karwacki Bode

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

January
1983

ACKNOWLEDGMENTS

The author gratefully acknowledges the services rendered by Drs. Jack A. Kavanagh and Samuel T. Mayo, the Director and other member of her thesis committee, the permission of Science Research Associates, Inc., for the use of their data for the study, the assistance of the SRA staff in the preparation of the manuscript, and the indulgence of her husband.

VITA

The author, Rita Karwacki Bode, is a native of Chicago and has obtained her education in the Chicago area. In June 1964, she received the degree of Bachelor of Arts with a major in psychology from DePaul University.

Upon graduation, she worked for the University of Chicago Industrial Relations Center and pursued a graduate degree in Industrial Psychology at DePaul University.

For the last 17 years, she has been in numerous research and development positions at Science Research Associates, Inc. and has been involved in scale and norms development, item tryouts, equating, and reliability-validity studies. She is currently a project director for high school testing programs.

While a graduate student at Loyola University, she made a presentation at the annual meeting of the American Educational Research Association (Los Angeles, 1981) on her thesis topic which was recently (October, 1981) referred to in an article in the American Psychologist Special Issue on Testing.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

Within the context of psychometric theory, studies of bias, whether in the tests themselves or in their use, are basically validity studies. In most cases, research on test bias consists of predictive studies in which scores on tests developed to or assumed to (in conjunction with other variables) maximize the prediction of some external criterion are correlated with future performance on that criterion. A test is considered biased if it under- or over-predicts the future success of the majority or minority group on the external criterion. Most research on item bias consists of construct-type studies using internal rather than external criteria; that is, performance on individual items in the test is compared to performance on other items in the test. In these studies, items are defined as biased if they do not measure the same construct for majority and minority groups.

Construct-type studies fall into two categories: a) item x group interaction studies using classical test theory in which significant interaction indicates that items are operating in different ways in different groups and, are hence, potentially biased; and b) item characteristic curve studies using a theoretical model that describes the characteristics of an item as a function of an underlying ability dimension in which unequal probability of success on an item for examinees of equal ability from different groups indicates bias.

Early item x group interaction studies used analysis of variance designs to determine whether significant interaction existed, and then examined performance on individual items to identify deviant items. One method used to identify items contributing to interaction, the delta plot method, compares estimates of item difficulty (proportion of examinees responding correctly to the item) to identify outliers (the term used to indicate biased items) from the main set of items; after adjustments are made for group differences, additional differences are considered a sign of bias. Another item x group interaction method, the item discrimination method (not covered in this study), uses the point biserial (item-test correlation) for the majority group as a standard against which the values for the minority group are compared; items with point biserials beyond the standard are identified as biased.

Item characteristic curve theory also uses item characteristics such as difficulty and discrimination, but defines them differently from the way they are defined in classical test theory. Difficulty is defined as the point on the ability continuum at which examinees have a 50/50 chance of answering the item correctly; discrimination is represented by the slope of the curve. An additional characteristic, the lower asymptote (the probability of a person of low ability guessing correctly on the item), is used. Variations of this three-parameter model, such as the one-parameter or Rasch model, make additional assumptions about the data, such as an equal discrimination level for all items and the non-existence of the guessing factor, but follow the same theoretical model. Bias is determined by calculating the area between curves for different groups (two- or three-parameter models), or statistically assessing the difference between the item parameters estimates from two different groups.

Statement of Problem

Of all the current methods for detecting biased items, the three-parameter item characteristic curve (ICC-3) is preferred theoretically because it provides the least confounded indices of both item difficulty and discrimination. More importantly, it is less likely to produce artifactual instances of bias due to true differences in group means since the parameter estimates are sample invariant. The three-parameter program (LOGIST) is expensive to run and requires a minimum of 40 items and 1000 examinees to reach stable parameter estimates. The problem thus is one of finding simpler methods that approximate three-parameter ICC results closely enough to recommend their use.

At present, the delta plot method is the most commonly used item bias detection technique because of its computational simplicity and its accepted use with smaller sample sizes. However, because of theoretical limitations (artifactual differences in item discrimination and differential guessing styles which appear as or obscure bias), the delta plot method will give good approximations of ICC-3 results only under certain conditions, those being, if the two groups being compared have true means that are very nearly equal and if all items are equally discriminating (Shepard, Camilli, & Averill, 1980). Concerns about the results obtained with the delta plot method when these conditions have not been met were first voiced by Lord (1977) and most recently by Linn, Levine, Hastings, and Wardrop (1981) in their studies using ICC-3.

In addition to theoretical limitation, another problem that affects all bias methods is the fact that there are no clear-cut decision rules for determining whether individual items are to be considered biased or unbiased. Methods that have been used to date are significance tests, identification of an arbitrary number of most biased items, and identification of abritrary cutoffs in the bias index.

According to Shepard, Camilli, & Averill,

> tests are unsatisfactory because they reflect how extreme a value is in the sampling distribution of the statistic rather than in the particular distribution of the item values obtained for a test; . . . the problem with identifying an arbitrary number of 'most biased' items is that it does not properly model our sense that biased items should be clearly discrepant from the pattern set by the other items in the test; . . . and the problem with arbitrary cutoffs is that two items with very similar indices can be considered as biased or unbiased simply because they are on either side of the cutoff.

Purpose of the Study

In a paper presented at the 1980 Johns Hopkins University Symposium on Educational Research entitled "Test Item Bias Methodology: State of the Art," W.H. Angoff suggested a number of procedural modifications to the delta method for use in item bias studies:

1. controlling for different ability levels in samples being compared using an external, if possible, criterion which is itself free of bias;

2. as recommended by Jensen (1980), using "pseudogroup" (majority ethnic groups whose average scores are similar to those of the minority group) performance compared to that of the total majority group as a baseline for interpreting majority-minority sample performance differences;

3. using baseline comparisons of samples from the same racial/ethnic group where differences exist on variables such as socioeconomic status or geographic area to estimate the variation within a majority-minority comparison that could be normally expected on a set of items;

4. replicating results of the analysis on comparable samples to determine the reliability of the method; and

5. analyzing biased items planted in the test to see if the technique detects the bias.

This study will address all but the last suggestion. Because the delta plot method may yield misleading results when groups under consideration score at widely different ability levels, and because all items do not have the same discrimination power, this study makes the assumption that adjusting for ability level differences in the majority versus minority and baseline comparisons should remove the effects of group ability level differences. Although not directly addressing differences in discrimination levels, this study assumes, as did the study of Sinnott (1980), that item deviation due to variance in discrimination level of the item would seldom produce extreme outliers but rather would only contribute to the general scatter of the plot.

Use of baseline comparisons to determine outlier identification, as found in the Sinnott study, takes into account the amount of scatter within the set of items which can differ from test to test. Therefore, it is assumed, that use of baseline comparisons will provide a more meaningful definition of outliers within the context of the particular set of items and may minimize the problem of variance in discrimination levels.

Use of several baseline analyses as compared to arbitrary cutoff values used in previous studies and a replication of results with each baseline should provide information on the consistency of results for a variety of baselines and, hopefully, lead to decisions on the most appropriate cutoff criterion to use in item bias studies incorporating the delta plot method.

CHAPTER II

REVIEW OF RELATED LITERATURE

The literature dealing with delta plot methodology can be categorized as follows: early studies in the development of the methodology; research on criteria to use for identifying outliers; research on the comparability of results using various item bias detection methods; and research on the consistency of results across comparable samples.

Early Studies in the Development

According to Angoff (1980), delta plot methodology goes back to the early days of psychometrics when L.L. Thurstone used it in connection with his Method of Absolute Scaling (1925). In this method, item difficulty (*p*) values are converted to normal deviates with a mean of 13 and a standard deviation of 4. These values, called deltas, are then plotted for two different groups on a bivariate graph, each pair of deltas for an item represented by a point on the graph. The plot of these points ordinarily appears in the form of an ellipse extending from lower left to upper right; for groups drawn from the same population, the scatterplot of points falls on this long narrow ellipse.

When groups differ in ability, the points still fall along the ellipse but are displaced vertically or horizontally. When groups differ in dispersion, the ellipse is tilted on an angle more or less steeply than 45 °. However, when the groups differ in type, the points for certain items fall outside the ellipse. When applied to item bias, the items falling at some distance from the ellipse may be regarded as contributing to item x group interaction.

Among the earliest research on item x group interaction were studies conducted by Cardall & Coffman (1963) and Cleary and Hilton (1968). The Cardall & Coffman study used a two-factor with repeated measures ANOVA design on three random samples each from rural white, urban white, and predominantly black samples; correlations between delta values were used to isolate the group or groups that contributed to the significant item x group interaction. The Cleary & Hilton study used a three-factor repeated measures ANOVA design on random samples of three socioeconomic groups within racial groups; bivariate plots of item sums (items were formula-scored) were used as indices of item x group interaction.

The first study to use analytic rather than graphic methods to define outliers was conducted by Angoff & Herring (1971). The procedures were later described by Angoff (1980) as follows:

"the formal procedure for measuring the departure of each item from the plot is to calculate its distance from the major axis of the ellipse. The equation for the major axis may be given in the form $y = ax + b$, where

$$a = \frac{(s_y^2 - s_x^2) \pm \sqrt{(s_y^2 - s_x^2)^2 + 4r_{xy}^2 s_x^2 s_y^2}}{2r_{xy}\, s_x s_y}$$

and

$$b = M_y - aM_x.$$

(It is recalled that the variables, x and y, are, respectively, the delta values for the two groups under consideration. Thus M_x and s_x, for example, denote the mean and standard deviation of deltas for the groups whose deltas are referred to the x-axis, and r_{xy} denotes the correlation between deltas for the two groups.) The formula for the distance, d_i, of each point, i, in the plot to the line (the major axis of the ellipse) is given as

$$d_i = \frac{ax_i - y_i + b}{a^2 + 1}."$$

The Angoff & Herring study was also the first to use within national or baseline comparisons to evaluate the results of a cross-national analysis. Inspectional rather than analytic methods were used to evaluate the baseline comparison in relation to the cross-national comparisons.

The study conducted by Angoff & Ford (1973) was the first to use matched samples in the analysis. The significant aspect of this study was the attempt to separate interaction due to racial differences from interaction due to ability differences by selecting and comparing performance of random and matched samples of students within each racial group. Among the results of this study was that, when matching was used, the between-race interaction decreased. Angoff (1975) later commented on this study saying,

> items x group interaction for the inter-race plot decreased, not quite to the level represented by the plots of random samples within race, but to a lower level nevertheless. . . . it would have dropped still further had we used a set of matching variables that were more highly correlated with the variables under study than the ones we did use (samples matched on math scores to analyze verbal items and vice versa).

A subsequent study by Angoff & Sharon (1974) used the analytic method for identifying outliers to summarize significant features of multiple-group comparisons when each group is compared to a "general" group.

A more recent study by Sinnott (1980), represents the first refinements since early use of the technique. These refinements consisted of: a) use of a "purified" criterion which eliminated items whose distance exceeded a specified amount from the calculation of the major axis, and b) use of values for the identification of outliers.

Research on Criteria for Outliers

A second category of research conducted subsequent to the development of the delta plot method dealt with studies in which various criteria for identifying outliers were used. In the period from 1975 to 1980, several studies used an arbitrary level in terms of the standard deviation of the distance of the item plots from the major axis. Strassberg-Rosenberg & Donlon, (1975, as reported in Carlton & Marco, 1980) and Donlon, Hicks, & Wallmark (1980) used a criterion of 1.5 standard deviations; Bleu & Ishizuka (1978, 1978, as reported in Carlton & Marco, 1980) used a criterion of 1.25 standard deviations; and Stern (1978, as reported in Carlton & Marco, 1980) used a criterion of three standard deviations. According to Donlon, Hicks & Wallmark, "such a level avoids undue capitalization on chance factors but should identify differences of practical significance." Humphreys (1979, as reported in Carlton & Marco, 1980) used graphic rather than distance measures to identify outliers. Scheuneman (1980a) used two criteria: an arbitrary number of "most biased" items (20) and an arbitrary cutoff (.75) using the distance formula presented on page 10. The most comprehensive study dealing with the determination of cutoffs for outlier identification was conducted by Sinnott (1980). As recommended by Angoff (1980), a baseline comparison was used to determine the point above which few items deviated from the line of best fit. The procedures used by

Sinnott follow the observations of Shepard, Camilli & Averill (1980). According to these researchers,

> biased items should be outliers. Outliers should be identified by gaps in the distribution of item values; these gaps could separate few or many items from the major cluster of item values . . . histograms of item bias indices (should be) inspected . . . in each data set. The 'most biased' items (should be) identified as those that (are) discrepant from the homogeneous and uninterrupted cluster of items.

Comparison of Detection Techniques

A third category of research conducted between 1978 and 1980 dealt with comparisons of various bias detection techniques. Among the major studies during this period were those conducted by Ironson & Subkoviak (1979); Rudner, Getson, & Knight (1980); and Shepard, Camilli, & Averill (1980). These studies have been summarized elsewhere (Devine & Raju, 1981) and will not be described here other than to note the basic conclusions that: a) the three-parameter item characteristic curve model was preferred; b) agreement among methods overall was reasonable; and c) the delta plot was the second-best method (next to the chi-square) in agreement with ICC-3 results. In a variation of these studies conducted by Subkoviak, Mack, & Ironson (1981) which dealt with another of Angoff's recommendations, ten intentionally biased items were added to the test and analyzed using four bias detection methods. In this study, the ICC-3 method was found to be the most effective in identifying the intentionally biased items, with the other methods (delta plot included) comparable to each other in detection ability.

Research on Consistency of Results

A fourth category of research dealt with consistency of results, that is, consistency of identification if items as biased across samples. Two studies, Scheuneman (1980a) and Bode (1981), found that while the consistency was better than chance, the results were less than expected. In the Scheuneman study, use of more than one bias detection method increased the consistency of identification and in the Bode study, sample size was found to be an important factor in consistency.

CHAPTER III

METHODOLOGY

Instruments

The items analyzed in this study came from the 1978 edition of the SRA Achievement Series, form 1, Level E Reading Vocabulary (40 items measuring literal and non-literal meanings); Math Concepts (30 items measuring whole numbers, fractions & decimals, geometry & measurement); and Language Arts Usage (40 items measuring verbs, pronouns & modifiers, clarity of expression, sentence structure, and sentence transformation) tests and the SRA Educational Ability Series (EAS), Level E (55 items measuring vocabulary, word grouping, numbers & series, and spatial) test.

The Technical Report #1 for this series (SRA, 1978) contains a description of the development of this instrument including the use of bias-free guidelines in item development, bias reviews of the items developed, pretesting, item selection including a statistical item bias study, use of content criteria to build the test forms, and tallies for fair representation (pp. 3-9).

Because this test was developed from an item pool from which items identified as biased were eliminated, it was necessary to select tests for this study that contained some at least marginally biased items. The basis for this selection was a previous study (Bode, 1981) of consistency of identification of biased items from the pretest and two subsequent samples. The vocabulary, math concepts, and language usage tests were selected because they contained at least two items identified as biased or marginally biased in the spring and fall reanalyses.

The EAS test was not analyzed for bias in this study but was instead used as an external control variable to adjust group performance by abilty level. In a previous analysis (unpublished portion of the above study) of the EAS items, no items were found to be biased toward either the majority or minority group.

Samples

Data available for this study consisted of item data for subsamples selected from two separate populations--the 1978 spring and fall standardization (norming) samples for this series. For each norming, "probability proportional to size" sampling was used to obtain a nationally representative sample; the combined number of students tested in the two normings was approximately 200,000 students across eights levels of the tests. Technical Report #1 (pp. 9-12, 18-30) and its addendum (SRA, 1979, pp. 2-5, 9-21) describe the samples, the sampling and norming procedures, and test characteristics for the norming samples; Technical Report #3 (SRA, 1980), among other things, describes the demographic characteristics of the standardization samples (pp. 20-33).

Representative samples were selected from the complete standardization samples in such a way that their Composite score distribution matched that of the complete standardization sample. Student data were sorted on a random variable and cases were pulled to meet the distributional requirements. Only students with complete test, sex, and racial/ethnic data were included in the samples.

Information available for selecting samples for this study included: identification of geographic region of the school district; size of the school district; demographic data for the school and community in which it was located; sex of the student; and racial/ethnic group membership of the student.

Procedures

From each representative sample, subsamples were drawn to create the majority (white), minority (black), and baseline (white) samples.

Majority/minority samples. The white and black samples were selected on the basis of the student-coded racial/ethnic group membership. For this study, data were combined by grade. The spring white sample consisted of 3845 students; the fall sample consisted of 2501 students. The spring black sample consisted of 698 students; the fall sample consisted of 750 students.

Geographic area-district size baseline samples. The geographic area-district size samples were selected from the white sample by categorizing school districts by geographic area (Northeast, North Central, South, and West) and distric size (small = districts with fewer than 3,000 students; large = districts with more than 50,000 students). Geographic area-district size samples with more than 300 students were considered of sufficient size for analysis. North Central-small and South-large had sufficient samples sizes for both spring and fall samples and were, therefore, selected for analysis. The spring North Central-small sample consisted of 338 students from 17 schools in 11 districts; the fall sample consisted of 323 students from 16 schools in 11 districts. The spring South-large sample consisted of 457 students from 38 schools in seven districts; the fall sample consisted of 309 students from 41 schools in seven districts.

SES baseline samples. The socioeconomic status (SES) high and low samples were selected from the white sample by estimating a composite (similar to that calculated by the Census Bureau as reported in Spiegelman, 1968) based on the occupation of the head of household (high = professional, executive, businesspeople; low = unskilled laborers, unemployed, migrant workers, etc.); family income (high = more than $25,000; low = less than $10,000); and educational level of the head of household (high = some college or more; low = less than high school). Schools were identified as high SES if at least two of the variables were rated as high with the third rated at least as average and as low SES if at least two of the variables were rated as low with the third rated no more than average. The spring high SES sample consisted of 986 students from 32 schools in 23 districts located in eight geographic regions; the fall sample consisted of 493 students from 31 schools in 20 districts located in seven geographic regions. The spring low SES sample consisted of 405 students from 24 schools in 20 districts located in seven geographic regions; the fall sample consisted of 275 students from 23 schools in 16 districts located in seven geographic regions.

Pseudogroup baseline samples. The pseudogroup baseline samples were selected by pulling samples of white students with the same EAS score distribution as that of the black sample. In order to obtain cell frequencies for the sample pull, frequency distributions of the EAS raw scores were obtained for the white and black samples. The spring pseudogroup sample consisted of 698 students; the fall sample consisted of 750 students.

Traditional item analysis (consisting of _p_ values and point biserials for each alternative) was performed on the white, black, and baseline samples for the vocabulary, concepts, usage, and EAS items. For the white, black, and baseline samples, except for the pseudogroup baseline samples, sample means on the EAS were compared to that of the national norm group to obtain the ratio of the national-to-group means for each sample. This ratio was then applied to the item _p_-values for each sample to create adjusted-for-ability item difficulty data.

Delta plot methodology was applied to the white versus black, white versus pseudogroup, South-large versus North Central-small, and high-SES versus low-SES comparisons using both adjusted and unadjusted data. For each comparison, deltas were calculated from the item difficulty data and pairs of deltas were plotted on bivariate graphs. Using the formulae described on page 10, the major axis and the distance of each item from the major axis were calculated.

For each baseline comparison, frequency distributions of the distance (d) values were obtained and, based on these distributions, outlier cutoffs were determined. Outliers were defined as the absolute value of the extreme distance values for the baseline samples which were characterized by gaps in the distribution that set them apart from the main cluster of items. To allow for varying dispersion across tests, outlier cutoffs were determined separately by test.

Statistical Analysis

In order to determine whether the use of baseline cutoffs for identifying outliers produced more consistent results than using arbitrary values, further analyses were made of the data from the delta plots. The outlier cutoffs for spring and fall baselines were first averaged so that the values used in both samples were the same. The baseline and arbitrary cutoffs were then used to identify items in the black-white comparison as outliers, separately for the spring and fall samples. Finally, each item was classified into low, moderate, and high bias indices and the classifications were compared for the spring and fall samples to determine the consistency of classification.

While the most obvious kind of comparison to make between the bias indices obtained for the two samples is a correlation, the use of this method suggests that even very low values of the bias indices are meaningful. According to Scheuneman (1980a),

> it is more likely that when the degree of bias in the item is low or non-existent, the indices reflect only random variation among responses in the groups being compared. Hence, there is no real reason to expect high agreement in indices except when bias exists at least in moderate degrees.

Instead, kappa coefficients were calculated, kappa being a procedure for comparing classifications on two different occasions which calls for the computation of the percent of agreement between two classifications beyond what would be expected by chance (Cohen, 1968). In order to compute kappa coefficients, the frequency distributions of distance values for each baseline were used to establish cutoffs to classify items as follows: those with absolute distance values greater than the outlier cutoffs for each baseline (characterized by a gap in the distribution) were classified as high bias; those in the portion of the distribution around which cell frequencies dropped off were classified as moderate bias; and those clustering around the zero values were classified as low bias. For the arbitrary values, items with distance values greater than .75 were classified as high bias; those with values between .40 and .75 were classified as moderate bias; and those with values less than .40 were classified as low bias.

Contingency tables were constructed consisting of the number of items classified accordingly in the spring and fall samples. The formula for the kappa coefficient is defined as

$$k = \frac{P_o - P_c}{1 - P_c}$$

where p_o = the obtained proportion of items classified in the same way (high, moderate, or low bias) in both sets of data (diagonal cells in the contingency table) and p_e = the proportion of items expected to be classified in the same way by chance (using the 3-way chi-square procedures).

Because the classifications of low/moderate/high bias could be ordered, weighted kappa could be used to take into account partial agreement. A weighted coefficient of agreement was computed by assigning weights to the different cells in the contingency table. Weights of one were used for perfect agreement (diagonal cells), zero weights were used for the high/low cells, and an intermediate weight of .5 was used for the high/moderate and low/moderate cells (Cicchetti & Fleiss, 1977). These weights were then multiplied by the corresponding entries in the chance and obtained proportion contingency tables. Weighted kappa, according to Cohen, is similarly defined as

$$k' = \frac{P_o' - P_c'}{1 - P_c'}$$

Weighted kappa coefficients vary from negative values for poorer than chance agreement through zero for chance agreement to plus one for perfect agreement.

The hypothesis that agreement is significantly better than chance was tested by calculating the standard error of weighted kappa, the critical ratio of weighted kappa to its standard error, and by referring the critical ratio to the standard normal distribution. Fleiss, Cohen, and Everitt (1969) found the large sample standard error of weighted kappa to be estimatable by

$$S.E._{k}' = \frac{1}{(1 - P_c')\sqrt{N}} \left[\sum_{i=1} \sum_{j=1} p_{i.} p_{.j} (w_{ij} - (\bar{w}_{i.} + \bar{w}_{.j})^2 - p_c'^2 \right]^{\frac{1}{2}}$$

The final analysis consists of the comparison of kappa and weighted kappa coefficients using the baseline versus arbitrary cutoff values. In addition to the comparisons using adjusted-for-ability data, the analyses were repeated using unadjusted data to determine whether the use of the adjustment improved the consistency of identification of outliers. The ranking of procedures in terms of producing the most consistent results, in addition to considerations such as sampling and selection "errors" in cutting criteria, were used to make recommendations on the appropriateness of procedure use.

CHAPTER IV

RESULTS

Because analysis of data for samples of comparable ability avoids one of the confounding factors in using the delta plot method, adjustment for ability was used in this study. The anchor test raw score means that were used to adjust data for each sample to resemble performance of an average group, the weights that were applied to item difficulty data to adjust for ability level differences, and the differences from the norm group means of each sample mean as reflected by the weights, are presented in Table 1. As can be seen from the data for both spring and fall, the white, large and small districts, and high SES samples scored above the national average and the black, pseudogroup, and low SES samples scored below. Greater differences existed between the white and black samples, and obviously between the white and pseudogroup samples, than between the high and low SES samples; small differences existed between the large and small district samples. In terms of consistency from spring to fall, weights were comparable for the white and low SES samples but higher in spring than fall for the remaining samples.

Table 1

Anchor Test Ability Adjustment Data

	Spring			Fall		
Sample	mean	weight	diff	mean	weight	diff
White	32.31	.941	-1.92	34.36	.944	-1.93
Black	21.56	1.410	8.83	25.23	1.285	7.20
Pseudogroup	21.70	--	8.69	25.21	--	7.22
S-Large	34.67	.877	-4.28	38.51	.842	-6.08
NC-Small	32.07	.948	-1.68	36.20	.896	-3.77
High SES	34.82	.873	-4.43	39.19	.828	-6.76
Low SES	27.48	1.106	2.91	30.13	1.076	2.30
Norm	30.39			32.43		

The means, standard deviations, and correlations between delta values for the black-white and baseline samples are presented in Table 2. Summaries are presented for analyses using the adjusted and unadjusted data. As can be seen from these data, when unadjusted, the items were usually easier for the white, large district, and high SES samples but, when adjusted, easier for the black, small district, and low SES samples. Pseudogroup sample data were not adjusted for ability because the samples were previously pulled to match the ability score distribution of the black samples. In terms of the consistency of delta values, the comparison group performance (black versus white samples, white versus pseudogroup samples, large versus small district samples, and high versus low SES samples) showed correlations of .90 or greater in all samples for the concepts test and in selected baselines for the vocabulary and usage tests. (Angoff (1975) assumed that unbiased tests would have delta correlations of .98 or above; obviously some bias still existed in these tests.) The pseudogroup baseline produced the highest correlations when compared to either adjusted or unadjusted baseline or black-white comparisons. In general, lower correlations were found for adjusted than for unadjusted vocabulary data but slightly higher correlations were found for unadjusted rather than adjusted concepts and usage data.

Table 2

Delta Value Summary Data

		Vocabulary		Concepts		Usage	
Sample		Fall	Spring	Fall	Spring	Fall	Spring
Unadjusted Data							
Black	mean	12.758	13.328	12.521	13.052	12.571	13.043
	s.d.	1.513	1.398	2.044	1.812	1.262	1.288
White	mean	10.549	10.924	11.075	11.372	10.990	11.102
	s.d.	1.468	1.556	1.883	1.767	1.226	1.331
r		.873	.878	.936	.944	.887	.845
White	mean	10.549	10.924	11.075	11.372	10.990	11.102
	s.d.	1.468	1.556	1.883	1.767	1.226	1.331
Pseudogroup	mean	12.053	12.579	12.251	12.741	12.543	12.815
	s.d.	1.447	1.447	1.807	1.654	1.156	1.309
r		.977	.967	.970	.971	.965	.954
S-Large	mean	9.646	10.427	10.224	10.908	9.956	10.557
	s.d.	1.596	1.668	1.872	1.930	1.245	1.479
NC-Small	mean	10.227	10.831	10.617	10.983	10.831	11.168
	s.d.	1.536	1.751	2.017	2.002	1.401	1.370
r		.891	.878	.933	.970	.940	.946
High SES	mean	9.540	10.167	10.276	10.880	9.794	10.576
	s.d.	1.628	1.798	1.840	1.725	1.339	1.429
Low SES	mean	11.461	11.927	11.803	12.221	11.826	11.980
	s.d.	1.417	1.624	1.905	1.789	1.265	1.321
r		.907	.901	.932	.945	.881	.891
Adjusted Data							
Black	mean	10.759	10.946	10.191	10.308	10.650	10.521
	s.d.	2.691	2.614	3.607	3.445	2.227	2.506
White	mean	10.073	11.423	11.561	11.840	11.442	11.574
	s.d.	1.276	1.371	1.682	1.588	1.074	1.181
r		.837	.850	.900	.915	.877	.841
S-Large	mean	11.321	11.570	11.733	11.948	11.477	11.638
	s.d.	1.019	1.248	1.306	1.512	.820	1.104
NC-Small	mean	11.227	11.304	11.577	11.448	11.696	11.576
	s.d.	1.157	1.539	1.616	1.774	1.120	1.245
r		.886	.882	.932	.971	.934	.951
High SES	mean	11.367	11.423	11.861	11.952	11.504	11.696
	s.d.	.993	1.313	1.282	1.354	.851	1.060
Low SES	mean	10.847	11.092	11.181	11.435	11.281	11.214
	s.d.	1.695	2.092	2.255	2.239	1.481	1.671
r		.884	.864	.907	.913	.857	.866

Because lower correlations indicate more dispersion or the existence of outliers, baselines with higher correlations corresponded with lower cutoffs and, therefore, more identified outliers. The outlier cutoffs established for each baseline and arbitrary values and the items identified as outliers using each of these criteria are presented in Table 3. As expected, in most of the analyses, the pseudogroup baseline had the lowest cutoff and highest number of identified outliers. The SES baseline produced the consistently highest cutoffs and, therefore, the fewest outliers. In all cases, the pseudogroup baseline cutoffs were lower than the arbitrary cutoffs and, in most cases, the district size and SES baseline cutoffs were higher than the arbitrary cutoffs.

In terms of adjusted versus unadjusted data, unadjusted data consistently had higher cutoffs and fewer outliers than adjusted data. In terms of consistency of items identified across spring and fall samples, about half of the vocabulary and usage items identified in either sample were identified in both but in concepts, the ratio was much less. Finally, in terms of consistency of items identified as outliers in both spring and fall samples as compared to the items identified as biased in the previous study (Vocabulary = 1, 5, 14; Concepts = 15, 29; and Usage = 15, 18, 39) as reported in Bode (1981), using the pseudo-

Table 3

Arbitrary and Baseline Cutoffs Using Adjusted And Unadjusted Data And Outliers Identified

		Fall	Spring
Baseline	cutoff	outliers	outliers
Vocabulary-Adjusted			
Pseudogroup	.65	1-5,9,13-14,16-18,21, 31,37,39-40	1,4-6,9,13-14,16-17, 21,31,37,39-40
Area/Size	.95	1,5,16	1,5,14,16,39-40
SES	.90	1,5,16,40	1,5,14,16,39-40
Arbitrary	.75	1,3-5,9,16,21,37,39-40	1,5,9,13-14,16-17, 21,31,37,39-40
Concepts-Adjusted			
Pseudogroup	.75	1-3,5,8,15,19,21,25-26	8,18,20,22,25,28
Area/Size	.70	1-3,5,8,15,19,21,25-26, 29	8,16,18,20,22,25,28
SES	1.05	5,15,25	8,18,20,22,25,28
Arbitrary	.75	1-3,5,8,15,19,21,25-26	8,18,20,22,25,28
Usage-Adjusted			
Pseudogroup	.60	1,5,7,15,18-20,22,39	3,5,7,10,15,18-20,39
Area/Size	.50	1,3,5,15,18-20,22,36, 39	3,5,7,10,13,15,17-20, 24,32,37,39
SES	1.00	18	5,7,10,15,18,20
Arbitrary	.57	1,5,7,15,18,20	5,7,10,15,18,20
Vocabulary-Unadjusted			
Pseudogroup	.65	1,2,5,9,14,16,18,22-23, 26,37	1,5,9,13,14,16, 23
Area/Size	.95	5,9,14	1,5,9,14
SES	1.25	5,14	14
Arbitrary	.75	5,9,14,16,18,23	1,5,9, 13-14,16
Concepts-Unadjusted			
Pseudogroup	.75	5,15,17,29	15,18,28,29
Area/Size	.85	15,17,29	18,29
SES	1.05	15,29	--
Arbitrary	.75	5,15,17,29	15,18,28,29
Usage-Unadjusted			
Pseudogroup	.60	15,18,22,39	5,7,10,15,18,20,39
Area/Size	.65	15,18,22,39	5,7,10 ,15,18,39
SES	1.05	18	15,18
Arbitrary	.75	15,18,22,39	10,15,18,39

group baseline, all vocabulary and usage items previously identified were identified as outliers, in concepts using adjusted data, neither of the previously identified items were identified as outliers, and in the remaining samples, partial agreement was found.

The amount of overlap in items identified as outliers between the spring and fall samples were verified by the results of the kappa and weighted kappa analyses. The coefficients using each baseline and arbitrary cutoffs, separately for adjusted and unadjusted data, are presented in Table 4. As expected, the coefficients for the concepts items were consistenly lower than those obtained for the vocabulary and usage items. In terms of the criteria (cutoff) which produced the greatest consistency, there was little or no consistency across tests (geographic area-district size analyses were more consistent in two of three tests using adjusted data and SES analyses were most consistent using unadjusted data). When looking at both adjusted and unadjusted results, one-third of the comparisons favor each baseline.

In terms of baseline use producing more consistent results than arbitrary values, in all comparisons at least one baseline was more consistent than the arbitrary values. Here again, there was an equal split in the number of analyses in which one or more of the baselines was more

Table 4

Kappa and Weighted Kappa Results

Baseline	kappa	weighted kappa	s.e.m.	critical ratio	conf. level
Vocabulary-Adjusted					
Pseudogroup	.586	.670	.218	3.071**	.9989
Area/Size	.629	.688	.343	2.003*	.9772
SES	.584	.644	.367	1.756*	.9608
Arbitrary	.434	.577	.236	2.451**	.9929
Concepts-Adjusted					
Pseudogroup	.087	.215	.348	.617	.7324
Area/Size	.158	.313	.278	1.267	.8980
SES	.160	.236	.717	.330	.6255
Arbitrary	.150	.289	.300	.966	.8340
Usage-Adjusted					
Pseudogroup	.468	.580	.365	1.588	.9441
Area/Size	.443	.482	.401	1.202	.8849
SES	.196	.344	.446	.711	.7611
Arbitrary	.406	.524	.412	1.272	.8980
Vocabulary-Unadjusted					
Pseudogroup	.395	.531	.294	1.810*	.9649
Area/Size	.605	.648	.457	1.418	.9207
SES	.729	.756	.569	1.328	.9082
Arbitrary	.461	.593	.365	1.625	.9474
Concepts-Unadjusted					
Pseudogroup	.389	.455	.369	1.232	.8907
Area/Size	.338	.348	.549	.634	.7357
SES	.063	.171	.901	.190	.5753
Arbitrary	.356	.373	.514	.725	.7673
Usage-Unadjusted					
Pseudogroup	.491	.589	.399	1.476	.9306
Area/Size	.476	.576	.511	1.127	.8708
SES	.624	.682	1.009	.676	.7517
Arbitrary	.579	.633	.598	1.198	.8810

* significant at .05
**significant at .01

consistent than arbitrary values. Finally, in terms of adjusted versus unadjusted data use, the general results show use of unadjusted data producing greater consistency than use of adjusted data in the majority of the cases. (In usage, unadjusted data produced more consistent results than adjusted in all baselines for both kappa and weighted kappa. In vocabulary, the pseudogroup and district-size baselines were more consistent using adjusted data and SES baseline and arbitrary values produced more consistent results using uadjusted data. In concepts, only SES baseline produced more consistent results using adjusted data; arbitrary values and the other baselines were more consistent using unadjusted data.)

Also found in Table 4 are the standard errors of weighted kappa, the critical ratios of weighted kappa to its standard errors, and confidence limits for each baseline and arbitrary value analyses. As seen by these data, the only coefficients which were significant at the .05 significance level or above were the vocabulary-adjusted analyses for all baselines and the vocabulary-unadjusted analysis for the pseudogroup baseline. In most cases, the standard error values for pseudogroup the the lowest of the baselines and lower than those for arbitrary values. In all cases, the values for SES were the highest of the baselines and higher than for arbitrary values.

CHAPTER V

DISCUSSION

Factors that need to be taken into account in setting the cutoff criteria for identifying outliers (whether it be based on arbitrary values or a baseline comparison) consist of 1) the consistency of each procedure in identifying the same items across comparable samples; 2) the impact of making "Type I" versus "Type II" errors of classification; and 3) the ease or difficulty in obtaining baseline samples.

This study addressed the first factor by looking at the consistency of identification of outliers using cutoffs based on arbitrary values versus those based on three baseline comparisons--pseudogroup, geographic area-district size, and socioeconomic status--and by comparing results obtained using adjusted-for-ability versus unadjusted black-white and baseline data.

Although the study did not produce consistent results in terms of one baseline producing more consistent results than the others or arbitrary values across all tests, several outcomes were of interest. First was the tendency of baselines that were more objective and easy to measure so as to produce the least error. The standard errors were the lowest for the pseudogroup baseline (sample pulled to match another distribution) and highest for the SES baseline (estimates of socioeconomic status of the students in the school with one standard criterion applied across all areas of the country). The high standard errors may have been due to the small size of the analysis samples (number of items in the tests). According to Fleiss & Cicchetti (1978),

> unless one's sample is very large (at least $16k^2$, where k is the number of categories in the scale), the standard error formula should be used with caution for setting confidence limits on the population values of weighted kappa.

Comparisons in which the SES baseline was the most consistent produced higher values of weighted kappa coefficients that those in which the pseudogroup baseline was most consistent with area/size somewhere in between. Because the SES baseline had the highest cutoffs, few items were identified as outliers. Of the items identified, one can be sure that they are indeed outliers, but cannot tell how many more undetected "true" outliers existed in the test. Using the more conservative psedogroup baseline, more items were identified as outliers. Of these items, one can be sure than they constitute all of the "true" outliers but can't tell if all of them are "true" outliers. The fact that the pseudogroup baseline identified the most outliers and SES identified the least has an interesting implication in breaking down black-white performance differences into confounding factors--controlling only for ability level differences accounts for a very small proportion of the black-white performance difference but using socioeconomic differences accounts for a large proportion of such differences.

The most surprising outcome of the study was the lack of improvement in the consistency using data that were adjusted for ability level differences. One would have assumed that the removal of one confounding factor in group differences would have enhanced the results of the delta method. Perhaps the correlations between the ability and test scores (vocabulary = .73 and .74 for fall and spring, respectively; concepts = .66 and .69 for fall and spring, respectively; and usage = .72 and .73 for fall and spring, respectively) indicate that a more highly correlated anchor score would be more effective. There are situations, however, in which using the adjustment factor is necessary. When an item bias study is conducted as part of a pretest study of items in a pool which are to be used to develop final forms, units of items are usually taken by distinct and separate samples, and therefore, adjustment for ability, either by using an external score common to all the samples or by imbedding anchor items in each pretest unit, are needed to "standardize" performance across samples within the same majority or minority group.

The most troublesome aspect of the study was that of classifying items as biased or unbiased by selecting as a cutoff the exact point that separates two classifications within a baseline distribution of distance values. In classifying items for the kappa and weighted kappa coefficients, the cutoffs between moderate and high bias indices were characterized by gaps in the distribution. Where a gap existed, where should the cutoff be set--at the value or interval preceding the gap, succeeding the gap, or somewhere in between? The original decision was made to use the midpoint between the distance values on either side of the gap. To see if different results would have been obtained had a different cutoff been used, a second set of classifications of items into the low, moderate, and high bias categories were made--this time with the cutoff between moderate and high bias set at the delta value preceeding the gap. For example, in the fall High versus Low SES vocabulary analysis, the three highest distance values were .83, 1.19, and 1.25. In the first analysis, the cutoff was set at the midpoint between .83 and 1.19; in the second analysis, it was set at .83 (values of .84 and above were considered high bias). Kappa and weighted kappa coefficients were calculated using these new classifications. In no case did the results of the second analysis indicate that using the second set of cutoffs would have produced more consistent results. In all but one analysis, no signficant change was found in the rankings of the baselines.

One factor that cannot be ignored in interpreting the results of this study is that the items used were previously screened for bias. One would have to assume that a bias study in which the delta values for the groups being compared had correlations greater than .90, the likelihood of finding many biased items or finding bias that would not be the result of random variation would be slim.

Evidence for the differences between a previously screened and unscreened item pool lies in a comparison of the delta correlations, as reported in SRA (1978), for the original item pool (vocabulary: 116 items and r = .687; concepts: 138 items and r = .831; usage: 130 items and r = .642) and for the final set of items (vocabulary: 40 items and r greater than .873; concepts: 30 items and r greater than .936; usage: 40 items and r greater than .845). In all instances the correlations were significantly higher in the final forms. Another limiting factor in the data used for this study was the relatively small item pool within each area (40 items in vocabulary and usage and 30 items in concepts). A pretest item pool conceivably contains 3-4 times as many items.

The second factor that needs to be taken into account in selecting a cutoff criterion is the impact of making "Type I" or "Type II" errors. As mentioned previously, when faced with the choice of selecting a low or high cutoff, such as when choosing between a pseudogroup or SES baseline, the number of outliers identified will differ significantly. If the most obviously biased items are to be eliminated from the item pool, using a high cutoff is sufficient; if items identified as outliers are to be reviewed for possible sources of bias or factors outside the item itself (such as the item reflecting valid performance differences that are important to the test or pecularities in the particular sample used), a low cutoff can be used and many items subjected to the review process. According to Scheuneman (1980b),

> the cutoff criteria . . . should be set with considered judgment taking into account the number of items which can reasonably be removed from the pool, the sample size used, and hence, the probable power of the procedure, the purpose of the exam, and the possible impact of either type of screening error. Where items are to be dropped automatically, more certainty may be desirable. Where items are to be reviewed, many may be tentatively identified as biased.

The third factor to take into account in selecting a cutoff criteria is the ease or difficulty of obtaining the baseline samples. For this study a relatively large and known population existed. Even so, data for the two grades in which the tests were administered had to be combined in order to obtain a sample of sufficient size. It would have been interesting to look at SES within geographic area-district size or SES within the black samples, but the samples obtained would have been too small for analysis.

In a pretest situation wherein each unit of items is taken by a distinct sample, difficulties in obtaining a spread of SES or geographic area-district size within each sample would be substantial. The easiest of the baselines to implement in a pretest situation would be the pseudogroup baseline. Pseudogroups only require that the majority sample have a sufficient ability range to allow for pulling a sample to match the minority group score distribution. This characteristic of a pretest sample is one that is not only needed for pulling a pseudogroup sample, but would also be desirable for any pretest sample.

CONCLUSIONS

A study using procedural modifications suggested for the delta plot method of item bias detection was conducted and, based on the replication in terms of the consistency of identification of items across comparable samples, the following conclusions were made:

1. adjustment for ability did not improve consistency of identification of outliers in the majority of cases;

2. use of the pseudogroup baseline produced the most consistent results in one-third of the analyses and was more consistent than arbitrary values in one-third of the analyses;

3. use of the SES and geographic area-district size baselines also produced the most consistent results each in one-third of the analyses and were more consistent than arbitrary values in one-third of the analyses.

Extenuating circumstances which might have affected these results include the high delta correlations for both baselines and white-black comparisons, the result of using items from a test built from a previously screened-for-bias item pool, and the relatively small number of items in the tests analyzed.

Prognosis for being able to adequately test the procedural modifications to the delta method investigated in this study--that is, a situation in which the extenuating circumstances noted above would not be contaminating factors--are not very good. The ideal situation would be one in which large samples of majority and minority students took previously unscreened-for-bias items. However, when large samples identified by majority and minority group exist, they usually are found for previously screened final test forms and when tests which have not previously been screened for bias exist, such as from pretest units, sample sizes are usually too small for analysis.

REFERENCES

Angoff, W.H. & Herring, C.L. Study of the appropriateness of the Law Score Admissions Test for Canadian and American students. Report No. LSAC-71-1. In Law School Admission Council, Reports of LSAC Sponsored Research: Volume II, 1970-74. Princeton NJ: Law School Admission Council, 1976.

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

Angoff, W.H. The investigation of test bias in the absence of an outside criterion. Paper presented at the NIE Conference on Test Bias, Annapolis, 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 Educational Research, "Test Item Bias Methodology: State of the Art," Washington DC, November 1980.

Bleu & Ishizuka. Unpublished study of SAT and TSWE items. In Carlton, S.T. and Marco, G.L. Methods used by ETS 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 DC, November 1980.

Bode, R.K. A comparison of the pretest and reanalysis results of an item bias study. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981.

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

Cicchetti, D.V. & Fleiss, J.L. Comparison of the Null Distributions of Weighted Kappa and the C Ordinal Statistic. Applied Psychological Measurement, Volume 1, No. 2, Spring 1977, 195-201.

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

Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 1968, 70, 213-220.

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 DC, March 1978.

Devine, P.J. & Raju, N.S. An investigation of the correspondence among four item bias identification methods. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981.

Donlon, T.F., Hicks, M.M., & Wallmark, M.M. Sex differences in item responses on the GRE. Applied Psychological Measurement, Volume 4, No. 1, Winter 1980, 9-20.

Fleiss, J.L. Cohen, J. & Everitt, B.S. Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 1969, 73, 323-327.

Fleiss, J.L. & Cicchetti, D.V. Inferences about weighted kappa in the non-null case. Applied Psychological Measurement, Volume 2, No. 1, Winter 1978, 113-117.

Humphreys, B.J. A review of data based on the performance of a sample of black students in southern colleges on the NTE Common Examinations. In Carlton, S.T. & Marco, G.L. Methods used by ETS 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 DC, November 1980.

Jensen, A.R. Bias in Mental Testing. New York: The Free Press, 1980.

Ironson, G.H. & Subkoviak, M.J. A comparison of several methods of assessing item bias. Journal of Educational Measurement, Volume 16, No. 4, Winter 1979, 209-225.

Linn, R., Levine, M.V., Hastings, C.N., & Washington, J.L. Item bias in a test of reading comprehension. Applied Psychological Measurement, Volume 5, No. 2, Spring 1981, 159-173.

Lord, F.M. A study of item bias using item characteristic curve theory. In N.H. Poortinga (Ed.) Basic Problems in Cross-Cultural Psychology, Amsterdam: Swits & Wiitlinger, 1977.

Rudner, L.M., Getson, P.R. & Knight, D.L. A monte carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, Volume 17, No. 1, Spring 1980, 1-11.

(a) Scheuneman, J.D. 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.

(b) Scheuneman, J.D. A posteriori analyses of biased items. Paper presented at the Johns Hopkins University Symposium on Educational Research, "Test Item Bias Methodology: State of the Art," Washington DC, November 1980.

Science Research Associates. SRA Achievement Series Technical Report #1. Chicago: SRA, 1978.

Science Research Associates. SRA Achievement Series Technical Report #1 Addendum. Chicago: SRA, 1979.

Science Research Associates. SRA Achievement Series Technical Report #3. Chicago: SRA, 1980.

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

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

Sinnott, L.T. Differences in item performance across groups. Research Report RR-80-19. Princeton NJ: ETS, 1980.

Speigelman, Mortimer. Introduction to Demography. Cambridge MA: Harvard University Press, 1968.

Strassberg-Rosenberg, B. & Donlon, T. Content influences on sex differences in performance on aptitude tests. In Carlton, S.T. & Marco, G.L. Methods used by ETS testing programs for detecting and eliminating item bias. Paper presented at the John Hopkins University Symposium on Educational Research, "Test Item Bias Methodology: State of the Art," Washington DC, November 1980.

Stern, J. Unpublished study of SAT/TSWE items. In Carlton, S.T. & Marco, G.L. Methods used by ETS testing programs for detecting and eliminating item bias. Paper presented at the John Hopkins University Symposium on Educational Research, "Test Item Bias Methodology: State of the Art," Washington DC, November 1980.

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.

Thurstone, L.L. A method of scaling psychological and educational tests. Journal of Educational Psychology, 1925, 16, 433-451.

APPROVAL SHEET

The thesis submitted by Rita Karwacki Bode has been read and approved by the following committee:

Dr. Jack A. Kavanagh, Director
Associate Dean, Loyola

Dr. Samuel T. Mayo
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. 9, 1982
Date

Jack A. Kavanagh
Director's Signature