A COMPARISON OF TWO DIFFERENTIAL ITEM FUNCTIONING DETECTION METHODS: LOGISTIC REGRESSION AND AN ANALYSIS OF VARIANCE APPROACH USING RASCH ESTIMATION

DISSERTATION

Presented to the Graduate Council of the
University of North Texas in Partial
Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

Marjorie Lee Threet Whitmore, B.A., M.A.
Denton, Texas
August, 1995

A COMPARISON OF TWO DIFFERENTIAL ITEM FUNCTIONING DETECTION METHODS: LOGISTIC REGRESSION AND AN ANALYSIS OF VARIANCE APPROACH USING RASCH ESTIMATION

DISSERTATION

Presented to the Graduate Council of the
University of North Texas in Partial
Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

Marjorie Lee Threet Whitmore, B.A., M.A.
Denton, Texas
August, 1995

Whitmore, Marjorie Lee Threet, A Comparison of Two Differential Item Functioning Detection Methods: Logistic Regression and an Analysis of Variance Approach Using Rasch Estimation. Doctor of Philosophy (Educational Research), August, 1995, 107 pp., 13 tables, bibliography, 75 titles.

Differential item functioning (DIF) detection rates were examined for the logistic regression and analysis of variance (ANOVA) DIF detection methods. The methods were applied to simulated data sets of varying test length (20, 40, and 60 items) and sample size (200, 400, and 600 examinees) for both equal and unequal underlying ability between groups as well as for both fixed and varying item discrimination parameters. Each test contained 5% uniform DIF items, 5% non-uniform DIF items, and 5% combination DIF (simultaneous uniform and non-uniform DIF) items. The factors were completely crossed, and each experiment was replicated 100 times.

For both methods and all DIF types, a test length of 20 was sufficient for satisfactory DIF detection. The detection rate increased significantly with sample size for each method.

With the ANOVA DIF method and uniform DIF, there was a difference in detection rates between discrimination parameter types, which favored varying discrimination and decreased with increased

sample size. The detection rate of non-uniform DIF using the ANOVA DIF method was higher with fixed discrimination parameters than with varying discrimination parameters when relative underlying ability was unequal. In the combination DIF case, there was a three-way interaction among the experimental factors discrimination type, relative ability, and sample size for both detection methods.

The error rate for the ANOVA DIF detection method decreased as test length increased and increased as sample size increased. For both methods, the error rate was slightly higher with varying discrimination parameters than with fixed. For logistic regression, the error rate increased with sample size when relative underlying ability was unequal between groups. The logistic regression method detected uniform and non-uniform DIF at a higher rate than the ANOVA DIF method. Because the type of DIF present in real data is rarely known, the logistic regression method is recommended for most cases.

TABLE OF CONTENTS

LIST OF TABLES

Table | Page

CHAPTER 1

INTRODUCTION

Overview of Differential Item Functioning Detection

A fundamental requirement of measurement is that test scores should be valid. A test should measure examinee ability accurately without regard to the subject's membership in any demographic group. Although sources of variation unrelated to the construct of interest cannot be eliminated entirely, efforts should be made to ensure that these sources of variation do not put any subpopulation at a disadvantage. If extraneous sources of variance are distributed differently for identifiable subgroups on a test item, the item is considered biased (Crocker & Algina, 1986). The presence of bias is a cause for concern because tests are used as gatekeepers for educational opportunities and employment advances. Legislative actions as well as lawsuits have resulted from perceptions that tests are biased against certain groups (Faggen, 1987; McAllister, 1993).

Bias in testing can be divided into two subcategories: selection bias and item bias. Studies of selection bias look at fair use of tests such as college admissions examinations and licensing examinations. Selection bias is studied through the comparisons of test scores to external criteria.

Item bias is studied by looking at the test structure itself through both judgmental and statistical methods. It is the statistical analyses of group differences on item characteristics that detect differential item functioning (DIF).

Although the phrases are sometimes used interchangeably, *differential item functioning* is more meaningful and neutral terminology than *item bias* when statistical properties are studied (Holland & Thayer, 1988; Humphreys, 1986). DIF detection refers to any empirical method used to flag items for possible item bias. Perhaps the relationship between item bias and DIF was best explained by Camilli (1993):

> An item is said to "function differently" for two or more groups if the probability of a correct answer to a test item is associated with group membership for examinees of comparable ability. Statistical indices of DIF are designed to identify such test items. If the degree of DIF is determined to be practically significant for an item and the DIF can be attributed plausibly to a feature of the item that is irrelevant to the test construct, the presence of this item on the test biases the ability estimates of some individuals. This compound condition, when satisfied, indicates item bias. (pp. 397-398)

To calculate statistical indices of DIF, a population is usually divided into comparison groups referred to as reference and focal groups, majority and minority groups, or base and comparison groups. Item statistics are then calculated for each group. If item statistics differ between groups after adjusting for ability differences, an item is said to exhibit DIF. Because these techniques use test scores as a measure of ability, only items biased relative to the test itself can be identified.

These procedures can be categorized into three broad classes: traditional classical test theory methods, chi-square methods, and latent trait theory methods. Of these, latent trait methods are theoretically preferred (Shepard, Camilli, & Williams, 1984), but they are computationally intense (expensive), usually need minimum sample sizes of 1,000, and often require test lengths of at least 40 items. In practice, the population of minority examinees may be too small to allow implementation of a latent trait method (Bleistein, 1986). Much research in the area of DIF detection focuses on alternatives that can be applied under less restrictive conditions and that require fewer computations.

The current method of choice, developed by Holland and Thayer (1988), is based on the Mantel-Haenszel chi-square statistic. The Mantel-Haenszel procedure is the DIF detection method used by many practitioners, including those at Educational Testing Service (Dorans & Holland, 1993). The technique's popularity is due to favorable performance comparisons with latent trait methods while taking substantially less time to calculate. The method is so well accepted that Holland and Wainer (1993) mentioned it as a standard against which new methods could be judged before adoption by measurement practitioners.

Rationale for Study

Although the Mantel-Haenszel procedure eliminates the time-consuming calculations inherent in latent trait DIF detection methods, it is not free from shortcomings. One troublesome condition is small sample size. Mazor, Clauser, and Hambleton (1991) found that with sample sizes of 500 the method's DIF detection rate was lower than 50%. Another situation in which the Mantel-Haenszel procedure performs poorly is in the presence of group differences in item discrimination, a condition known as *non-uniform DIF*. This problem has been illustrated analytically (Hambleton & Rogers, 1989) and empirically (Rogers, 1989; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990).

To overcome these difficulties, two procedures have been proposed as alternatives to the Mantel-Haenszel procedure. Both have been shown through simulation studies to outperform the Mantel-Haenszel technique. The ANOVA method using Rasch-based estimates (Tang, 1994) has been found to be more powerful than the Mantel-Haenszel method with small sample sizes. The second method, based on logistic regression, has been shown to detect non-uniform DIF, which was not found by the Mantel-Haenszel method (Rogers, 1989; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). Unlike the Mantel-Haenszel procedure, both of these methods treat ability as a

continuous variable and thus utilize more information (Linacre & Wright, 1987; Swaminathan & Rogers, 1990).

Two studies of logistic regression as a DIF detection method have used simulated data to compare the performance of logistic regression to the Mantel-Haenszel technique. Under all studied conditions, logistic regression has been superior to the Mantel-Haenszel method, most notably in the presence of non-uniform DIF (Rogers, 1989; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). A third study (Tian, Pang, & Boss, 1994) used a variation of the Swaminathan and Rogers' technique on real data and found that the logistic regression method outperformed the Mantel-Haenszel procedure. The Tang (1994) study used simulated data to compare the performance of the ANOVA DIF detection method with Rasch-based estimates to the Mantel-Haenszel technique. In all simulated situations, the ANOVA DIF detection method was superior to the Mantel-Haenszel method, most notably in small sample sizes. Because the ANOVA DIF detection and logistic regression methods have been show to be preferable to the Mantel-Haenszel method in these studies, a study to directly compare these two procedures is warranted. What remains unknown is the relative performance of these two DIF detection methods under identical conditions. No research has been done directly comparing the two methods.

The ANOVA DIF detection method using Rasch-based estimates has been applied only to the detection of uniform DIF. The existence of non-uniform DIF violates an assumption of the Rasch model. However, Rudner, Getson, and Knight (1980) detected non-uniform DIF with a Rasch-based method by interpreting the lack of fit between the data and the model as DIF. Swaminathan and Rogers (1990), by contrast, are among the researchers who assert that Rasch-based DIF methods will not detect non-uniform DIF. Angoff (1993) further stated that Rasch-based DIF detection methods are likely to find artifactual DIF in cases of model-data misfit. There seems to be no empirical verification of these claims.

The ANOVA DIF detection method has been limited to the study of test lengths of 30 items. Sample sizes have ranged from 200 to 1,200 total examinees. Wright and Stone (1979) recommended a minimum of 200 total examinees and a minimum test length of 20 when using the Rasch model. Examination of the ANOVA DIF detection method has also been confined to data in which discrimination was constant for all items.

Logistic regression has not been evaluated for use when sample sizes were less than 500 total examinees or when text lengths were less than 40 items. Furthermore, the logistic regression method has been used only with reference and focal groups constrained to have equivalent underlying ability distributions. Rogers (1989) proposed further study

when the reference and focal groups have unequal ability distributions, when sample sizes are as small as 100 per group, or when test lengths are as short as 20 items. Consequently, these two methods were compared on identical sets of data with known DIF and related factors known to affect DIF detection.

Research Questions

The following questions were examined with respect to the logistic regression and ANOVA methods of DIF detection:

1. Do significant interactions or main effects exist for test length, sample size, discrimination types, and relative ability when the type of DIF is uniform?

2. Do significant interactions or main effects exist for test length, sample size, discrimination type, and relative ability when the type of DIF is non-uniform?

3. Do significant interactions or main effects exist for test length, sample size, discrimination type, and relative ability when both uniform DIF and non-uniform DIF are present?

4. Do significant interactions or main effects exist for test length, sample size, discrimination type, and relative ability for false positive errors?

Delimitations

The generated data simulated responses to dichotomously scored test items, with no guessing and no omitted answers. The factors under study were delimited to include only three sample sizes (200, 400, and 600 examinees), three test lengths (20, 40, and 60 items), two relative ability distributions (equal and differing by one standard deviation), and two item discrimination assumptions (constant and varying). The use of simulated data in this study permitted knowledge about which items actually contained DIF and were anticipated to be detected by DIF detection methods. However, artificially created data may not necessarily reflect DIF as it actually occurs with real examinees in any given test situation (Subkoviak, Mack, Ironson, & Craig, 1984).

Definitions

DIF, an acronym for differential item functioning, is present in a test item if the probability of a correct answer differs between equally able members of separate demographic groups.

Uniform DIF occurs when an item is uniformly more difficult for one group than another across all ability levels. The case in which only item difficulty varied between groups was considered uniform DIF.

Non-uniform DIF occurs when the difference in difficulty between groups varies across ability levels. The case in which only item discrimination varied between groups was considered non-uniform DIF.

Combination DIF exists when both item difficulty and item discrimination vary between groups.

Focal group is the group of examinees against whom DIF is generated.

Reference group is the group of examinees against whom no DIF is generated.

Latent Trait Theory encompasses models that related unobservable abilities (latent traits) to test item performance. An in-depth discussion of this theory can be found in Hambleton (1989).

Item characteristic curve represents the probability of a correct response to an item as a function of the ability measured by the test that contains the item.

Item discrimination (a-parameter) is proportional to the slope of the item characteristic curve at the point of inflection.

Item difficulty (b-parameter) is the location of the point of inflection on the item characteristic curve ability scale.

Guessing (c-parameter) is the lower asymptote of the item characteristic curve.

Examinee ability estimate (θ) is the examinee latent trait or ability level.

Rasch model is a latent trait model which assumes that item discrimination is equal for all items and that no guessing occurs.

Observed score is the actual item score.

Expected score is the score expected for an item based on Rasch estimation.

Residual score is the difference between an observed and an expected score.

CHAPTER 2

REVIEW OF THE LITERATURE

Methods to detect the presence of differential item functioning (DIF) can be placed in three broad categories: traditional (classical true-score) methods, chi-square methods, and latent trait theory methods. The categories differ in their approach to detecting DIF. The various methods in each category are described below.

Traditional Differential Item Functioning Methods

Traditional methods are those based on classical true score theory. The item difficulty and item discrimination used in this measurement theory are analogous, but not equivalent, to those in latent trait theory. Classical true score theory, as well as the DIF detection methods based on this theory, yields uncomplicated calculations that are easy to compute. The major problem is that these methods are test and sample dependent. The results do not generalize to the general population because they depend on the particular group of examinees and the specific test items used to obtain the item statistics. A more detailed discussion of classical true score theory weakness is contained in Crocker and Algina's (1986) work.

Traditional Analysis of Variance

The application of analysis of variance (ANOVA) to the study of DIF originated with Cleary and Hilton (1968). A number of designs have been used, such as Ethnicity x Gender x Age x Items (Jensen, 1974). Usually, the factorial designs include independent grouping variables such as gender, with item score as the dependent variable (Osterlind, 1983). The magnitude of the item-by-group interaction in relation to other sources of variance is used as an indicator of DIF (Jensen, 1973). In the presence of significant item-by-group interaction, post hoc multiple comparison procedures are used to detect specific items with DIF (Plake, 1981; Plake & Hoover, 1979).

The ANOVA method is easily understood, relies on simple calculations, and does not require a large number of examinees for implementation. Unfortunately, highly discriminating items, easy items, and difficult items will falsely indicate DIF if the groups differ in achievement level (Camilli & Shepard, 1987).

Following publication of Camilli and Shepard's (1987) denouncement, use of ANOVA methods for DIF detection seemed to end. However, ANOVA was successfully used with matched pairs of examinees in a split-plot factorial design (Seong, 1990).

Transformed Item Difficulties

When the transformed item difficulties, or delta plot method is used, the proportion of correct responses for an item, known as *p*-values,

is calculated separately for each group. These *p*-values are transformed into *z*-values corresponding to the (1-*p*)th percentile of the standardized normal distribution. Delta values are calculated using the formula $\Delta = 4z + 13$, then plotted on a coordinate system whose axes represent the comparison groups. An ellipse is fitted to the points, and the distance between each point and the major axis of the ellipse is computed as a measure of DIF for the item represented by the point (Angoff, 1972). Items are ranked by the magnitude of this difference, with large differences indicating more DIF than small differences. Baghi and Ferrara (1989) found very high agreement between Rasch and delta plot DIF indices.

The transformed item difficulties approach is simple and inexpensive to implement; nevertheless, this method tends to confound item difficulty with item discrimination; hence, when utilized with groups of differing mean ability, highly discriminating items will show spurious DIF (Angoff, 1982; Shepard, Camilli, & Williams, 1985). A modification suggested by Angoff has been found to be inadequate (Seong & Subkoviak, 1987; Shepard et al., 1985). However, Shepard et al. partialled point biserials out of the existing delta indices with good results and recommended further study.

Standardization Method

The first DIF detection method, based on classical test theory, that considered differences in ability between groups was the standardization method (Dorans & Kulick, 1983). Empirical item characteristic curves are formed for each group using scaled scores for the *x*-axis and *p*-values for the *y*-axis. The differences in *p*-values between groups at each score level are weighted and summed. The resulting value is then compared to a cut-off value. Because there are no test statistics associated with the technique, practitioners refer to it as a DIF *description* method rather than a DIF detection method (Dorans & Holland, 1993).

The standardization method was developed at Educational Testing Service to detect DIF in the *Scholastic Aptitude Test*. Large sample sizes are required for stability (Dorans & Kulick, 1983) making the method impractical for many situations. Outside the Educational Testing Service, the standardization method has been applied in at least one study (Masters, 1988). No empirical verification using data with known DIF seems to have been conducted.

Partial Correlation

The partial correlation index, proposed by Stricker (1982), is the correlation between group membership and success on an item after partialling out scores on the test with the item omitted. The phi coefficient was used to determine significance.

Researchers who have used the partial correlation index cite its ease of computation, applicability to small samples, and ability to handle disproportionate sample sizes (Willson, Nolan, Reynolds, & Kamphaus, 1989). A sample size of at least 1,500, with no fewer than 300 examinees in the smallest subgroup, was recommended for stability (Stricker, 1984).

Regression Bias

Raju and Normand (1985) advocated the development of a regression line for each group, using *p*-values as the criterion and total test score as a predictor variable. Non-identical regression lines imply that an item has DIF. The *F*-ratio for equality of two regression lines was used to indicate significant DIF. Raju and Normand claimed that this method was easy to implement and could be used with sample sizes as small as 150. When compared to other DIF detection methods, this procedure's correlation was strongest with chi-square methods and weakest with latent trait methods. Raju and Normand acknowledged that logistic regression was theoretically preferred to the regression bias method, in spite of more complex computations.

Conclusion

With the exception of the standardization method, none of these methods seems to have found acceptance with measurement practitioners. This is understandable in light of the test and sample

dependence inherent in any procedure based on classical test theory. The advantage of simple calculation becomes smaller as computers become faster, more powerful, and less expensive.

Chi-Square Differential Item Functioning Methods

The general strategy of chi-square methods is to partition each group into subgroups based on test scores, then to compare proportions of correct responses within each level. DIF is assumed to exist if these proportions are unequal between groups. Variations are created by specifying different null and alternative hypotheses.

General Chi-Square Methods

Scheuneman (1979) developed a DIF index similar to χ^2, which was subsequently found not to be distributed as the χ^2 model required, and, therefore, was no longer recommended (Osterlind, 1983). Other χ^2 DIF detection methods have since been proposed.

Shepard and Camilli (1981) reported a modification by Camilli to the Scheuneman procedure that resulted in an index with a χ^2 distribution. Camilli's χ^2 requires creation of three to five ability intervals, followed by calculation of χ^2 for each interval. The χ^2 values for each interval are summed, and the result is tested for significance. Ironson (1982) also indicated how this statistic can easily be calculated

with a 2 x 2 contingency table for each ability level (right/wrong by reference/focal groups).

Two other χ^2 methods for identifying DIF have been presented by Marascuilo and Slaughter (1981). One of these methods tests the null hypothesis of no group differences in proportion correct at any ability level against the alternative hypothesis of a constant group difference across ability levels. Another method is based on partitioning χ^2 across the number of ability levels, using planned pairwise comparisons.

Although these methods are easy to understand, use sample sizes as small as 200, and can be tested for significance, there are some disadvantages. The magnitude of the chi-square statistic could change if the score interval cutoffs are changed. If the score distributions between the groups are different, chi-square can become inflated. Finally, by treating the continuous variable test score as a discrete variable, information can be lost (Ironson, 1982).

Logit Models

Mellenbergh (1982) proposed the creation of three-way tables (score level x group x response), followed by tests for a main effect of score level or comparison group, and a score by group interaction. If a model including only a constant and a score group fit the data, the item was free of DIF.

An interative model was developed by Van der Flier, Mellenbergh, Adèr, and Wijn (1984). Essentially, at the *i*th iteration, the *i* items with the most DIF were removed, and the observed score indicator of ability for the $(i + 1)$th iteration was computed from the remaining items (Intraprasert, 1986).

Mantel-Haenszel

The Mantel-Haenszel procedure was developed by Holland and Thayer (1988) and has been used by many researchers. Examinees are divided into ability intervals based on test score. For each interval, a 2 x 2 table (item score by group membership) is constructed for the item to be studied. The null hypothesis is that the odds of an item being correct is the same in each group across all ability levels. It is based on the logit model reduced by the removal of the interaction parameter, which results in a more powerful test when there is only uniform DIF, but which prevents the detection of non-uniform DIF (Rogers, 1989).

The Mantel-Haenszel procedure assumes that the underlying ability distributions of the reference and focal groups are equal (Linacre & Wright, 1987; Zwick, 1990), a condition which may not exist in practice. Camilli and Smith (1990) noted that the power of the Mantel-Haenszel method depended on the total sample size at any particular raw score. Hambleton and Rogers (1989) demonstrated analytically that the Mantel-Haenszel statistic was not designed to detect

group differences in item discrimination, a condition known as non-uniform DIF.

Using artificial data, researchers have found that the Mantel-Haenszel procedure performs poorly with poor discrimination in the DIF-containing items, with small differences in between-group item difficulties (Mazor, Clauser, & Hambleton, 1991), with very difficult DIF-containing items (Clauser, Mazor, & Hambleton, 1991; Mazor et al., 1991), and in the presence of non-uniform DIF (Rogers, 1989; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). Ryan (1991) found that obtaining stable estimates with the Mantel-Haenszel procedure required larger sample sizes than the Educational Testing Service recommended minimum total sample size of 500. Mazor et al. (1991) also found that fewer than half of the DIF items were detected with sample sizes of 500.

Logistic Regression

The use of logistic regression to detect DIF has been explored by Rogers (1989), Swaminathan and Rogers (1990), Rogers and Swaminathan (1993), and Tian, Pang, and Boss (1994). This method is a generalization of the Mantel-Haenszel procedure (Wainer, 1993). The probability of an individual's correct response to an item is given by the formula

$$\pi(\vec{x}) = \frac{e^{z}}{1+e^{z}}$$

where

$$z = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g).$$

The parameter τ_0 is the intercept, τ_1 is the coefficient for ability, τ_2 corresponds to group difference in item performance, τ_3 corresponds to group by ability interaction, θ is the observed ability of the individual, and g represents group membership (e.g., $g = 1$ for examinees in the reference group and $g = 0$ for examinees in the focal group). The parameters τ_0, τ_1, τ_2, and τ_3 can be estimated by the method of maximum likelihood, as explained by Hosmer and Lemeshow (1989).

An item is considered free of DIF if $\tau_2 = \tau_3 = 0$. A likelihood ratio test and a Wald test statistic have been used in DIF research to determine the significance of these coefficients. Swaminathan and Rogers (1990) and Rogers and Swaminathan (1993) used the Wald statistic, but Rogers (1989) used both the Wald statistic and the likelihood ratio test and found no differences in the results (H. J. Rogers, personal communication, December 2, 1993). Hosmer and Lemeshow (1989) recommended the likelihood ratio test for theoretical and practical reasons, noting that the likelihood ratio test was easier to implement using existing software packages.

The likelihood ratio test compares the logistic regression model that includes all parameters to a model with the restriction $\tau_2 = \tau_3 = 0$. These models can be expressed as

$$z_{FULL} = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3 (\theta g)$$

and

$$z_{REDUCED} = \tau_0 + \tau_1\theta.$$

The values of the parameters are estimated for each model. Each model's log likelihood L is then calculated using

$$L = \ln(\prod_i \zeta(\vec{x}_i))$$

where

$$\zeta(\vec{x}_i) = \pi(\vec{x}_i)^{y_i}[1-\pi(\vec{x}_i)]^{1-y_i}$$

and $y_1, \ldots, y_n$ are the item scores of the n examinees.

Finally, the test statistic is

$$G = -2(L_{REDUCED} - L_{FULL})$$

which has a chi-square distribution with 2 degrees of freedom.

Programs that estimate logistic regression parameters and calculate the corresponding log likelihoods are available in commonly used statistical software packages. In terms of computer time, Rogers and Swaminathan (1993) found that the logistic regression method was three to four times as expensive as the Mantel-Haenszel method, yet much less expensive than item response theory methods.

The logistic regression procedure's detection rate for DIF is affected by model-data fit, sample size, test length, and type of DIF (Rogers, 1989; Rogers & Swaminathan, 1993). With non-uniform DIF, Swaminathan and Rogers (1990) found detection rates of 50%, using 500 total examinees and test lengths of 40, 60, and 80 items. They achieved a 75% detection rate using 1,000 examinees and an 80-item

test. With non-uniform DIF, they detected at least 75% of the DIF items with 500 examinees and 100% of the items with 1,000 examinees.

Conclusion

Of all these methods, the Mantel-Haenszel is the most widely accepted due to its ability to approximate results of latent trait methods. The newest method, logistic regression, has great promise due to the ability to detect non-uniform DIF.

Latent Trait Differential Item Functioning Methods

Latent trait theory encompasses both item response theory and the Rasch measurement model. The latent trait methods have an advantage over the previous methods because they yield sample-free estimates (Hambleton, Swaminathan, & Rogers, 1991). With the notable exceptions of the one-parameter item response theory and Rasch models, large samples are required, and the calculations are time-consuming. The Rasch measurement model uses only item difficulty and person ability in calculating the probability of success on an item. It is appropriate for small sample sizes and reflects consistent, sufficient, and efficient estimates (Wright & Stone, 1979).

Item Response Theory Methods

Item response theory models relate the probability of success on an item to an examinee's ability and to a combination of the item's

difficulty, discrimination, and probability of being answered correctly by guessing. When difficulty, discrimination, and guessing are all incorporated into the item response theory model, it is a three-parameter model. If the probability of guessing an answer correctly is zero, the model is a two-parameter model. If all items are equally discriminating and if the chance of guessing the correct answer to any item is zero, the model is a one-parameter model.

Item response theory comparison methods. The detection of DIF by comparing item parameters or item characteristic curves has been addressed by numerous authors (Bleistein, 1986; Lord, 1980; Ironson, 1982; Hambleton et al., 1991; Shepard, Camilli, & Williams, 1984; Thissen, Steinberg, & Wainer, 1993). Comparison groups are formed, and separate parameters are calculated for the groups. An item is free of DIF if the parameters are equal between the groups or if the area between the item characteristic curves is zero. There are problems associated with these methods. The test of significance for the parameter comparison method (Lord, 1980), as well as formulas for the area between item characteristic curves and the associated tests of significance (Raju, 1988, 1990), are rather complicated in the two-parameter and the three-parameter case. Hambleton et al. noted that parameter comparison in the three-parameter case may not be very powerful and that minimum sample size for the significance test was unknown. There is no test of

significance for the area between item characteristic curves when the *c* parameters are unequal because the area is infinite (Raju, 1988).

Pseudo-item response theory method. Because item response theory comparison methods based on a two- or three-parameter model require large sample sizes, their use has been precluded when groups were small. One method, termed *pseudo-item response theory* by Shepard et al. (1985), was proposed by Linn and Harnisch (1981) for use when the number of examinees in a minority group was too small to utilize a traditional item response theory approach. Item response theory item parameters are obtained for all examinees in the combined group, and the ability scale is divided into quintiles. The difference between expected probability of a correct response and observed item response is computed for members of the group of interest, and the average difference for each interval is computed. The index of DIF is the sum of these differences.

Shepard et al. (1985) stated that this approach was the "method of choice" (p. 103) for DIF detection with small sample sizes. However, Seong and Subkoviak (1987) found that this procedure was slightly less accurate than a simpler chi-square approach. Scheuneman (1990) noted that when the mean scores for the groups are quite different, this method is unlikely to detect differences in guessing ability between groups.

Rasch Methods

The Rasch measurement model describes the relationship of item difficulty and person ability to the probability of a correct answer by expressing the probability of a correct answer with the ratio

$$P(\theta) = \frac{e^{(\beta_v - \delta_i)}}{[1+e^{(\beta_v - \delta_i)}]}$$

where β_v is the ability of person v, and δ_i is the difficulty of item i. Person abilities and item difficulties can be estimated from item scores by application of the unconditional method (UCON), as explained by Wright and Stone (1979). Computer source code for programs that use the UCON algorithm has been made available by both Linacre (1990) and Baker (1992). The estimation process is known as *data calibration*.

Small sample sizes are not as problematic for the Rasch model as for other latent trait models (Lord, 1980), and the calculations are much simpler. The formula for the area between item characteristic curves reduces to the absolute value of the differences in the difficulty parameter (Hambleton et al., 1991); thus, DIF detection by calculating the area between curves is equivalent to DIF detection by parameter comparison.

DIF detection based on the Rasch model has been shown to be theoretically preferable to the Mantel-Haenszel procedure. The Rasch model and the Mantel-Haenszel method require the same assumptions. However, the Rasch model treats test scores as a continuous variable,

whereas the Mantel-Haenszel procedure uses test scores as a blocking variable, resulting in a loss of data (Linacre & Wright, 1987). Schulz, Perlman, Rice, and Wright (in press) noted that at some test score levels, the Mantel-Haenszel procedure can have incomplete 2 x 2 tables that cannot be used. This loss of data reduces reliability.

Separate calibration t-test approach. Wright and Stone (1979) discussed a graphical method equivalent to a procedure that Smith (1993) referred to as the separate calibration *t*-test approach. This appeared to be the most widely used Rasch DIF detection method. Item difficulty parameters are computed separately for the reference and focal group. For each item, the difference in difficulty between the reference and focal groups is divided by the square root of the difference of the standard errors. If the result exceeds the cutoff value of ±2, the presence of DIF is indicated. With empirical data, Englehard, Anderson, and Gabrielson (1989) found this method to be more reliable than the Mantel-Haenszel procedure. Schulz et al. (in press) found an almost perfect correlation between Rasch and Mantel-Haenszel techniques when the comparison groups had equal achievement. This approach was more sensitive to DIF and was more reliable in small focal groups (100 to 200 examinees) than the Mantel-Haenszel procedure.

Goodness of fit approach. A second type of Rasch-based method compares the model-data fit of the entire population studied to the

model-data fit of the focal group (Wright, Mead, & Draba, 1976). A study using simulated data found that this procedure could not detect DIF in difficult items, but that it could detect non-uniform DIF (Rudner, Getson, & Knight, 1980). Whereas researchers including Swaminathan and Rogers (1990) and Angoff (1993) stated that the lack of a discrimination parameter in the Rasch model rendered it useless for the detection of non-uniform DIF, Rudner et al. suggested that the utilization of goodness-of-fit by this method allowed for the poor fit of items that differ in discrimination between groups to be detected as DIF.

A modified version of this method utilizes a single calibration for the combined reference of focal groups, allowing for smaller total sample sizes. When compared to the separate calibration *t*-test approach, it had a Type I error rate. Otherwise, there seemed to be no real differences in the results (Smith, 1993).

Analysis of variance method using Rasch estimation. Tang (1994), who developed this method, referred to it as item response theory ANOVA, and it can theoretically be used with item response theory models. For this study, the procedure used a Rasch model basis and was therefore referred to as the ANOVA method using Rasch estimation. When this technique is employed, the item difficulties and person abilities are estimated from the combined reference of focal groups. This use of a single calibration allows total sample size to be as small as 200 without

violating the minimum sample size recommendation of Wright and Stone (1979). Expected scores are computed for each item for each examinee. For dichotomous items, an examinee's expected score is equal to the probability of correctly answering the item. For each item, each person's expected score is subtracted from the observed score to obtain a residual score. Analysis of variance is performed, with residual scores as the dependent variable and group membership as the independent variable. The test statistic is the *F*-ratio.

To date, only one study (Tang, 1994) has used this method. With simulated data, this method was more powerful than the Mantel-Haenszel procedure when total sample size was smaller than 600. When ability was unequal between the groups and the DIF favored the focal group, the Mantel-Haenszel method was more powerful. Otherwise, the methods had comparable detection rates. The error rate for the ANOVA DIF detection method was higher when the mean ability level of the reference and focal groups differed.

Conclusion

In general, latent trait methods are more powerful than traditional (classical true-score) or chi-square methods of DIF detection. The major disadvantages of the latent trait methods are the requirement of large sample sizes and the complicated calculations. For small sample sizes, the ANOVA DIF detection method had comparable DIF detection rates.

CHAPTER 3

METHODS AND PROCEDURES

Simulated sets of raw score data for dichotomously scored items were generated for further analysis. A 3 x 3 x 2 design with a total of 36 experiments was used. Three test lengths (20, 40, and 60 items), three sample sizes (100, 200, and 300 persons per group), two mean ability relationships between groups (equal and unequal), and two item discrimination assumptions (constant and varying) were completely crossed. Both the analysis of variance (ANOVA) and logistic regression differential item functioning (DIF) detection methods were applied to this data. Each experiment was replicated 100 times. The experiments are explicitly numbered in Table 1.

Differential Item Functioning Types

The logistic regression DIF detection method has been examined with respect to uniform, non-uniform, and combination DIF, whereas the ANOVA DIF method has been studied only in the context of uniform DIF. The detection rates of the methods under study were examined with respect to all three types of DIF, as well as false positive errors.

Table 1

Definition of Experiments

Experiment	Discrimination	Ability	Items	People
1	Fixed	Equal	20	200
2	Fixed	Equal	20	400
3	Fixed	Equal	20	600
4	Fixed	Equal	40	200
5	Fixed	Equal	40	400
6	Fixed	Equal	40	600
7	Fixed	Equal	60	200
8	Fixed	Equal	60	400
9	Fixed	Equal	60	600
10	Fixed	Unequal	20	200
11	Fixed	Unequal	20	400
12	Fixed	Unequal	20	600
13	Fixed	Unequal	40	200
14	Fixed	Unequal	40	400
15	Fixed	Unequal	40	600
16	Fixed	Unequal	60	200
17	Fixed	Unequal	60	400
18	Fixed	Unequal	60	600
19	Varying	Equal	20	200

(*table continues*)

Experiment	Discrimination	Ability	Items	People
20	Varying	Equal	20	400
21	Varying	Equal	20	600
22	Varying	Equal	40	200
23	Varying	Equal	40	400
24	Varying	Equal	40	600
25	Varying	Equal	60	200
26	Varying	Equal	60	400
27	Varying	Equal	60	600
28	Varying	Unequal	20	200
29	Varying	Unequal	20	400
30	Varying	Unequal	20	600
31	Varying	Unequal	40	200
32	Varying	Unequal	40	400
33	Varying	Unequal	40	600
34	Varying	Unequal	60	200
35	Varying	Unequal	60	400
36	Varying	Unequal	60	600

Uniform Differential Item Functioning

Although the two DIF detection methods in this study have not been directly compared in previous studies, each had been compared to a

third method, the Mantel-Haenszel DIF detection method, with respect to uniform DIF detection. Swaminathan and Rogers (1990) and Rogers and Swaminathan (1993) had found that the logistic regression method was as powerful as the Mantel-Haenszel method in detecting uniform DIF and that the detection rate increased with sample size. Manipulating test length had no effect on the detection rate. Relative underlying ability of comparison groups and item discrimination type were not included in the studies. The ANOVA DIF detection method had been found to be more powerful than the Mantel-Haenszel method at detecting uniform DIF with the groups' sizes used in the present study, although the data used were limited to items with a constant discrimination parameter (Tang, 1994). Group ability differences had no clear effect on the detection rate, and the effect of test length was not studied. When the logistic regression and ANOVA DIF detection methods were compared directly, it was expected that both methods would perform satisfactorily and that, with constant discrimination, perhaps the ANOVA DIF detection method would have a higher DIF detection rate than the logistic regression method.

Non-Uniform Differential Item Functioning

In an analysis of empirical data, Hambleton and Rogers (1989) found that non-uniform DIF exists, yet cannot be reliably detected by all DIF detection methods. The logistic regression method had been shown

to detect non-uniform DIF, with higher detection rates associated with increased sample sizes (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Tian, Pang, & Boss, 1994). Rogers and Swaminathan also found that the method had a higher detection rate, with a test length of 80 items compared to a test with only 40 items. The effect of equal or unequal underlying ability of the comparison groups on the detection rates had not been studied, nor had the effects of constant versus varying discrimination parameters. The ANOVA DIF detection method's performance on non-uniform DIF had not been examined. Based on the work of Linacre and Wright (1987), a Rasch-based DIF detection method such as the ANOVA DIF detection method was not expected to detect non-uniform DIF.

Combination Differential Item Functioning

Rogers and Swaminathan (1993) referred to this type of DIF as "mixed" non-uniform. With the logistic regression method, they have found this type of DIF to be detected at or above the rate of uniform DIF and at a substantially higher rate than strictly non-uniform DIF. The detection rate of combination DIF had not been assessed for the ANOVA DIF method, and it was unknown whether the presence of non-uniform DIF in an item that already exhibited uniform DIF would enhance, degrade, or have no effect on the detection rate. None of the

experimental factors in the current study had been examined with respect to combination DIF detection.

False Positive Errors

Clearly, the utility of a DIF detection method would be compromised in spite of a high detection rate if it also frequently detected DIF where none actually existed. Swaminathan and Rogers (1990), as well as Rogers and Swaminathan (1993), found the logistic regression method to have a higher false positive error rate than the Mantel-Haenszel procedure, although the effects of the manipulated factors in the current study had not been directly investigated. Tang (1994) studied the ANOVA DIF's false positive error rate on items with constant discrimination and found an increase in the presence of group ability difference and increased sample size. Similar results were expected for the present study.

Experimental Factors

The experimental factors manipulated in this study have been shown to affect DIF detection. In general, smaller sample sizes, shorter test lengths, and unequal ability distributions make DIF more difficult to detect.

Test Length

Test length is an important consideration because longer tests result in more reliable measures of ability, everything else being equal. Rogers and Swaminathan (1993) found no significant difference in DIF detection rates using logistic regression to find uniform DIF with test lengths of 40 and 80. There was a small difference when the DIF was non-uniform. No studies have considered the impact of varying test length on the ANOVA DIF detection procedure. Wright and Stone (1979), however, have recommended a minimum test length of 20 items for data calibration using the Rasch model.

Sample Size

The performance of a DIF detection method with small sample sizes is of interest because the available pool of focal group members may be small. Swaminathan and Rogers (1990) found that the power of the logistic regression method increased with a sample size increase from 500 to 1,000 total examinees. Rogers and Swaminathan (1993) found an approximate 15% increase in uniform DIF detection rates and a 19% non-uniform detection rate increase when total sample size increased from 500 to 1,000 examinees. Tian et al. (1994) found a substantial increase in power with increased sample size when using the logistic regression method. Tang (1994) found that the ANOVA DIF method was more powerful under a variety of conditions when sample size

increased from 200 to 1,200 in increments of 200. Tang also found that the false positive error rate increased with sample size in the presence of group ability difference. Wright and Stone (1979) recommended a minimum sample size of 200 examinees for the Rasch model.

Relative Underlying Ability

Rogers (1989) suggested that the logistic regression method be studied when the reference and focal groups have unequal mean ability levels, noting that, in an actual testing situation, a focal group may have a lower mean ability than the reference group. Tang (1994) found the ANOVA DIF detection method to have a false positive error rate increase when the studied groups were unequal in ability and sample size increased. Schulz, Perlman, Rice, & Wright (in press) noted that Rasch-based DIF detection methods may be unable to detect non-uniform DIF when reference and focal groups differ in underlying ability.

Type of Item Discrimination

The Rasch model specifies that item discrimination be held constant. Varying the discrimination parameter will simulate misfit in the Rasch model, but not for the logistic regression procedure. The Rasch measurement model is generally robust to varying item discrimination (Baker, 1992), but the presence of non-uniform DIF may be obscured. In cases of model-data misfit, Rasch-based DIF detection

methods may find artifactual DIF (Angoff, 1993). The ANOVA DIF detection method using Rasch-based estimates has not been evaluated for use with varying item discrimination.

Construction of Simulated Data Sets

The data sets were constructed using an author-written BASIC computer program, the code for which is in Appendix A. Each data set contained 15% DIF items, a worst-case scenario, according to Rogers (1989). For each set, 5% of the items were more difficult for the focal group than for the reference group (uniform DIF), 5% of the items were more discriminating for the focal group than for the reference group (non-uniform DIF), and 5% of the items were more difficult and more discriminating for the focal group than for the reference group (combination DIF). The data were generated using the following formula:

$$P_i(\Theta) = (1+e^{-a_i(\Theta - b_i)})^{-1}$$

where $i = 1, 2, 3, \ldots, n$, for an n-item test, $P_i(\Theta)$ is the probability that a person of ability Θ answers item i correctly, Θ is the ability parameter, a_i is the discrimination parameter for item i, and b_i is item difficulty for item i.

The value of $P_i(\Theta)$ was calculated from the values of ability, item difficulty, and item discrimination, obtained as described below. This $P_i(\Theta)$ value was compared to a random uniform deviate from the interval

(0, 1). If the deviate was larger than $P_i(\Theta)$, an incorrect answer for the item was simulated; otherwise, a correct response was simulated.

Ability Parameter Simulation

For the equal ability case, abilities were normally distributed with a mean of 0 and a standard deviation of 1. In the unequal ability case, the reference group abilities were normally distributed with a mean of 0 and a standard deviation of 1, and the focal group abilities were normally distributed with a mean of -1 and a standard deviation of 1. These values approximate findings from actual data and have been used by several researchers in the study of DIF detection (Clauser, Mazor, & Hambleton, 1991; Donoghue & Allen, 1993; Mazor, Clauser, & Hambleton, 1991; Zwick, Donoghue, & Grima, 1993).

Discrimination Parameter Simulation

Item discrimination was equal to 1 for all items in the constant discrimination case, which reduced the equation to a Rasch model. In the case where item discrimination varied, the values were given by $(1.7)e^z$, where z is normally distributed with a mean of -.065 and standard deviation 0.13. Zwick et al. (1993) used item discrimination values obtained from this distribution, citing the close approximation to actual values. Other studies have used this type of approximation as well (Donoghue & Allen, 1993; Miller & Oshima, 1992).

Difficulty Parameter Simulation

Item difficulties were normally distributed with a mean of 0 and a standard deviation of 1. Studies have used these values based on their proximity to real data (Donoghue & Allen, 1993; Miller & Oshima, 1992; Rudner, Getson, & Knight, 1980).

Differential Item Functioning Simulation

To simulate both uniform and non-uniform DIF, item parameters were chosen so that the area between the item characteristic curves of the reference and focal group was equal to 0.6. In Tang's (1994) study of the ANOVA DIF method, items with simulated uniform DIF met this condition. Swaminathan and Rogers (1990) included items with simulated uniform and non-uniform DIF, which fulfilled this specification. Uniform DIF was simulated by adding .6 to the difficulty parameter for the focal group. Non-uniform DIF was simulated by setting the discrimination parameter equal to 1 for the reference group and 1.763073 for the focal group. In both cases, the area between the item characteristic curves was 0.6. Combination DIF was simulated by changing both the difficulty and the discrimination parameters as described herein. The remaining items, which were not manipulated to simulate DIF, were nevertheless subjected to both detection methods to test false positive detection rates.

Creation of Simulated Data Sets

A control file was written for each of the 36 experimental conditions. Each file resulted in 100 sets of raw score data. For example, the control file for the first experimental condition consisted of the following line: "c:\items\exp01dat" 20 100 1 100 1 1.

The parameters include the name of the output file ("c:\items\exp01dat"), the number of test items (20), the number of examinees per comparison group (100 per group), the filename extension of the first replicated data set (1), the filename extension for the last replicated data set (100), the choice of fixed discrimination (1), and the choice of equally able comparison groups (1).

When the data generation program was run with this control file as input, 100 sets of data were generated and written to the files

c:\items\exp01dat.1

c:\items\exp01dat.2

c:\items\exp01dat.3

(filenames for data sets 4 through 98)

c:\items\exp01dat.099

c:\items\exp01dat.100.

The data generation program was run with each of the 36 control files.

Analysis of Variance Differential Item Functioning Procedure

A BASIC computer program based on code by Linacre (1990) and adapted by the author for this study was used to compute the Rasch-based residual scores. A second program was written by the author to perform analysis of variance, with residual scores as the dependent variable and group membership as the independent variable. A significant *F*-ratio at the 0.05 level indicated the presence of DIF. The second program also counted the number of DIF items detected and the number of false positive errors. These programs are included in Appendix B and Appendix C, respectively.

Rasch residual score calculations. A control file was written for each of the 36 experimental conditions, which took each of the 100 replicated data sets for the experimental condition and computed the Rasch residual score for each item for each examinee. For the data sets created for experiment 1 (described in the preceding section), the file contained the following lines:

```
20      200      11
"c:\items\exp01dat.1"
"c:\items\exp01res.1"
"c:\items\exp01dat.2"
"c:\items\exp01res.2"
```

"c:\items\exp01dat.3"

"c:\items\exp01res.3"

(instructions for replications 4 through 98)

"c:\items\exp01dat.99"

"c:\items\exp01res.99"

"c:\items\exp01dat.100"

"c:\items\exp01res.100."

The first line included the number of items (20), the number of total examinees (200), and the column where the data began (11). The remainder of the lines contained the input and output filenames for each replication.

When the Rasch residual score calculation program was run with this control file as input, a set of residual scores was calculated for each set of simulated data. This program was run with each of the 36 control files.

Analysis of variance differential item functioning detection calculations. A control file was written for each of the 36 experimental conditions. Each control file instructed the ANOVA program to perform ANOVA with the residual scores obtained above as the dependent variable and group membership as the independent variable. For the residual scores obtained from experiment 1, the control file contents were

```
20      20      "c:\results\anova01.out"
:c:\items\exp01res.1"
"c:\items\exp01res.2"
"c:\items\exp01res.3"
(instructions for replications 4 through 98)
"c:\items\exp01res.99"
"c:\items\exp01res.100."
```

The first line included the test length (20), the total number of examinees (200), and the file to which the results were to be written ("c:\results\anova01.out"). The remaining lines were the filenames of residual score sets that were to be analyzed.

When the ANOVA program was run with this file as input, ANOVA DIF detection was performed on all the items in the 100 replications. The ANOVA program was run with each of the 36 control files.

Logistic Regression Differential Item Functioning Procedure

The Statistical Package for Social Sciences (SPSS) program was used to compute the logistic regression parameters and resulting log likelihood statistic necessary for the logistic regression DIF procedure. For each item, logistic regression was performed on the full model, using item score as the dependent variable and total score, group membership, and total score by group membership interaction as independent

variables. The reduced model, which had item score as the dependent variable and total score as the independent variable, was also evaluated. The log likelihood for the full model was subtracted from the log likelihood for the reduced model and the result multiplied by -2. The resulting statistic had a chi-square distribution with 2 degrees of freedom. When this statistic was significant at the 0.05 level, the presence of DIF was indicated. This program by the author is included in Appendix D.

The simulated data sets were analyzed using the first program listed in Appendix D. This program produced an extremely large amount of additional output that was not required for DIF detection; thus, a second program was written that edited the original output to a more manageable size and counted the number of DIF items detected and false positive errors. This second program follows the first in Appendix D.

Assessment of the Procedures

The data were analyzed with a separate four-way ANOVA procedure for each of the following: uniform DIF, non-uniform DIF, combination DIF, and false positive errors. The dependent variable was the percentage of items detected, and the independent variables were the experimental factors test length, sample size, relative ability, and discrimination type. The Statistical Analysis System (SAS) program used for this analysis is in Appendix E. The ANOVA *F*-test and/or

subsequent simple effects tests were used to determine significant differences in the experimental factors for the research questions posed.

CHAPTER 4

RESULTS

The data were analyzed using a four-way analysis of variance (ANOVA) procedure for each of the hypotheses under uniform differential item functioning (DIF), non-uniform DIF, and combination DIF conditions, as well as false positive errors. The dependent variable was the percentage of items detected, and the independent variables were the experimental factors test length, sample size, relative ability, and discrimination type. Any significant interactions were subject to further analysis.

In addition to the four-way ANOVA procedures, the mean percentage detection rates of the experimental factors were calculated. The mean percentage detection rates are presented in Table 2. The summary data results for the 36 experiments are in Appendix F.

Uniform Differential Item Functioning Detection

As can be seen from Table 3, when the ANOVA DIF method was used, sample size, discrimination type, and relative ability were all significant main effects. A significant interaction between sample size and discrimination type was found, as shown in Table 3. This

Table 2

Mean Percentage Detection Rates Over 100 Replications

Factor and level	Uniform DIF	Non-uniform DIF	Combination DIF	False positive errors
		ANOVA DIF method		
Test length				
20	63.92	28.50	79.00	7.03
40	65.17	29.46	77.75	6.47
60	63.08	29.72	77.58	6.30
Sample size				
200	42.49	19.28	63.60	5.61
400	70.29	29.18	83.11	6.68
600	79.39	39.22	87.63	7.51
Discrimination type				
Fixed	54.37	32.11	83.32	6.38
Varying	73.74	26.34	72.90	6.83
Relative ability				
Equal	70.57	29.97	74.96	6.42
Unequal	57.54	28.48	81.26	6.78
		Logistic regression DIF method		
Test length				
20	57.17	55.50	86.06	7.24
40	57.46	59.00	85.54	7.07
60	56.50	60.11	85.53	6.77
Sample size				
200	35.58	35.15	71.78	6.48
400	62.69	60.96	89.89	6.87
600	72.85	78.50	95.49	7.72

(*table continues*)

Factor and level	Uniform DIF	Non-uniform DIF	Combination DIF	False positive errors
Discrimination type				
Fixed	45.14	56.34	85.28	6.36
Varying	68.94	60.06	86.16	7.69
Relative ability				
Equal	63.09	57.17	83.48	6.26
Unequal	50.99	59.24	87.95	7.79

Note. DIF = differential item functioning; ANOVA = analysis of variance.

interaction rather than the contributing main effects was therefore investigated. The cell means for this interaction are presented in Table 4. It was found that, at sample sizes of 200 and 400, the detection rate for items with varying discrimination was over 20% greater than for items with fixed discrimination. At sample size 600, this difference was around 12%. Regarding the significant main effect of relative ability, the detection rate for items when ability was equal between reference and focal groups was higher than for items for which underlying ability differed between groups (Table 2).

As in the case with the ANOVA method, the factors of sample size, discrimination type, and relative ability were significant using the logistic regression method as shown in Table 3. The detection rate increased

Table 3

Analysis of Variance of the Effects of Experimental Factors on Uniform Differential Item Functioning Detection Rates

	ANOVA		Logistic regression	
Experimental factor	*F*	*p*	*F*	*p*
Test length (TL)	1.13	0.32	0.23	0.79
Sample size (SS)	379.68	0.00	356.74	0.00
Discrimination type (DT)	289.07	0.00	408.56	0.00
Relative ability (RA)	130.94	0.00	105.59	0.00
SS x DT	8.54	0.00	1.91	1.15
DT x RA	3.56	0.06	1.16	0.28
TL x RA	0.92	0.40	0.93	0.40
TL x DT	1.34	0.26	0.13	0.88
TL x SS	1.07	0.37	2.04	0.09
SS x RA	2.39	0.09	0.46	0.63
TL x SS x DT	1.33	0.26	0.91	0.46
TL x SS x RA	1.73	0.14	0.53	0.71
TL x DT x RA	3.40	0.03	2.02	0.13
SS x DT x RA	3.02	0.05	4.40	0.01
TL x SS x DT x RA	1.69	0.15	0.67	0.64

Note. ANOVA = analysis of variance.

Table 4

Interaction Cell Means for Analysis of Variance Differential Item Functioning Method and Uniform Differential Item Functioning

	Discrimination type		
Sample size	Fixed	Varying	*F*
200	31.31	53.67	128.41*
400	58.78	81.81	136.18*
600	73.03	85.75	41.57

*$p < .001$.

with sample size (Table 2). The detection rate was also higher for varying discrimination than for fixed. And, as with the ANOVA method, the detection rate was higher for equal ability than for unequal ability.

Non-Uniform Differential Item Functioning Detection

A significant interaction between discrimination type and relative ability is shown in Table 5 for the ANOVA method. This interaction was examined before main effects were considered, and the cell means are presented in Table 6. Although relative ability was not a significant main effect, it was found that, when ability was unequal, the detection rate for fixed discrimination items was over 12% greater than for

Table 5

Analysis of Variance of the Effects of Experimental Factors on Non-Uniform Differential Item Functioning Detection Rates

	ANOVA		Logistic regression	
Experimental factor	*F*	*p*	*F*	*p*
Test length (TL)	0.42	0.66	5.36	0.01
Sample size (SS)	100.17	0.00	440.44	0.00
Discrimination type (DT)	25.14	0.00	9.63	0.00
Relative ability (RA)	1.68	0.20	2.99	0.08
SS x DT	2.64	0.07	0.69	0.50
DT x RA	34.42	0.00	0.02	0.89
TL x RA	2.37	0.09	0.36	0.70
TL x DT	0.92	0.40	2.01	0.13
TL x SS	0.61	0.66	0.48	0.75
SS x RA	2.82	0.06	1.64	0.20
TL x SS x DT	0.96	0.43	2.43	0.05
TL x SS x RA	1.12	0.35	2.56	0.04
TL x DT x RA	2.31	0.10	0.89	0.41
SS x DT x RA	1.79	0.17	0.34	0.71
TL x SS x DT x RA	0.99	0.41	0.85	0.49

Note. ANOVA = analysis of variance.

Table 6

Interaction Cell Means for Analysis of Variance Differential Item Functioning Method and Non-Uniform Differential Item Functioning

	Discrimination type		
Relative ability	Fixed	Varying	*F*
Equal	29.48	30.46	0.36
Unequal	34.74	22.22	59.19*

*p .001.

varying discrimination items. When ability was equal, the difference in detection rates between discrimination types was less than 1%. Sample size was a significant main effect, and, as shown in Table 2, the detection rate increased with larger sample sizes. In all cases, the detection rate of non-uniform DIF using the ANOVA DIF method was extremely low.

When the logistic regression method was employed, as shown in Table 5, there were no significant interactions, and sample size and discrimination type were both significant factors. As indicated in Table 2, the detection rate increased as sample size increased. Varying discrimination items were detected at a significantly higher rate than

those with fixed discrimination (Table 5); however, the actual detection rates (Table 2) differed by less than 4%.

Combination Differential Item Functioning Detection

For the ANOVA DIF method, the significant three-way interaction among sample size, discrimination, and relative ability is denoted in Table 7. The discrimination by relative ability interaction was examined at each level of sample size. The discrimination and ability factors were found to interact at every level of sample size, with the smallest *p*-value at sample size 200, and the largest *p*-value at sample size 600 (i.e., the significance of the interaction decreased as sample size increased). The main effect of ability was then assessed for each level of discrimination type for each sample size. As can be seen in Table 8, the detection rates with equal ability are quite similar regardless of discrimination type at each level of sample size. In contrast, the detection rates with unequal ability differ by almost 33% between discrimination types at sample size 200 and by around 14% at sample sizes 400 and 600. With fixed discrimination, unequal ability had a much higher detection rate than did equal ability for all sample sizes, although the difference became less pronounced as sample size increased. With varying discrimination, there was less than a 3% difference in detection rates between equal and unequal relative ability except for sample size 200, where there was a difference of 12%. The detection rate increased as sample size

Table 7

Analysis of Variance of the Effects of Experimental Factors on Combination Differential Item Functioning Detection Rates

	ANOVA		Logistic regression	
Experimental factor	*F*	*p*	*F*	*p*
Test length (TL)	0.78	0.46	0.17	0.84
Sample size (SS)	211.40	0.00	267.24	0.00
Discrimination type (DT)	105.68	0.00	1.01	0.32
Relative ability (RA)	38.54	0.00	26.10	0.00
SS x DT	8.65	0.00	0.39	0.68
DT x RA	95.43	0.00	2.62	0.11
TL x RA	0.66	0.51	2.97	0.05
TL x DT	4.44	0.01	1.38	0.25
TL x SS	0.95	0.44	0.73	0.57
SS x RA	1.01	0.36	3.78	0.02
TL x SS x DT	1.32	0.26	2.94	0.02
TL x SS x RA	0.21	0.93	0.65	0.63
TL x DT x RA	0.32	0.73	1.75	0.17
SS x DT x RA	11.34	0.00	5.62	0.00
TL x SS x DT x RA	1.27	0.28	2.23	0.06

Note. ANOVA = analysis of variance.

Table 8

Interaction Cell Means for Analysis of Variance Differential Item Functioning Method and Combination Differential Item Functioning

Sample size	Equal ability	Unequal ability	*F*
200			
Fixed discrimination	61.17	82.22	71.84*
Varying discrimination	61.50	49.50	23.33*
400			
Fixed discrimination	78.17	94.28	42.06*
Varying discrimination	80.00	80.00	0.00
600			
Fixed discrimination	86.33	97.78	21.22*
Varying discrimination	86.21	83.78	0.22

*p .001.

increased across all combinations of discrimination type and relative ability. Items with fixed discrimination had a statistically equal or higher detection rate than those with varying discrimination across all combinations of relative ability and sample size.

As with the ANOVA method, the detection rate of the logistic regression method increased with sample size. Likewise, the same three-way interaction among sample size, discrimination type, and relative ability was present, although with a smaller (less significant) *F*-value. The interaction is indicated in Table 7.

An analysis of the interaction between discrimination type and relative ability at each level of sample size revealed a significant interaction at sample size 200. The cell means for this interaction are presented in Table 9. With fixed discrimination, items with unequal underlying ability between groups were detected at a higher rate than those with equal ability. With varying discrimination, items with equal underlying ability were detected at a slightly higher rate than those with unequal relative ability, although the difference was insignificant.

Table 9

Interaction Cell Means for Logistic Regression Method and Combination Differential Item Functioning at Sample Size 200

	Relative ability		
Discrimination type	Equal	Unequal	*F*
Fixed	68.33	75.33	10.66*
Varying	73.61	69.83	3.10

*$p < .01$.

As no interaction between discrimination type and relative ability existed at sample size 400 or 600, each of these experimental factors was analyzed individually for each sample size. Although discrimination type was not found to be significant, items with unequal underlying ability

were detected at a significantly higher rate than those with equal underlying ability. These cell means are presented in Table 10.

Table 10

Interaction Cell Means for Logistic Regression Method and Combination Differential Item Functioning

	Relative ability		
Sample size	Equal	Unequal	*F*
400	86.04	93.64	24.47*
600	93.34	97.64	8.06**

*p .001. **p < .01.

False Positive Errors

As shown in Table 11, test length, sample size, and discrimination type were all significant main effects for the ANOVA DIF method. The error rates for test lengths 40 and 60 were not significantly different, but for a test of 20 items, the error rate increased. The error rate increased as sample size increased, and it was slightly higher with varying discrimination than fixed discrimination.

With logistic regression, discrimination type was significant (Table 11), while varying discrimination had a higher error rate than fixed

Table 11

Analysis of Variance of the Effects of Experimental Factors on False Positive Error Detection Rates

	ANOVA		Logistic regression	
Experimental factor	*F*	*p*	*F*	*p*
Test length (TL)	8.39	0.00	2.94	0.05
Sample size (SS)	51.44	0.00	21.19	0.00
Discrimination type (DT)	8.62	0.00	69.04	0.00
Relative ability (RA)	5.51	0.02	91.41	0.00
SS x DT	0.99	0.37	1.55	0.21
DT x RA	0.18	0.67	7.60	0.01
TL x RA	0.04	0.96	1.71	0.18
TL x DT	0.20	0.82	0.53	0.59
TL x SS	0.72	0.58	0.19	0.95
SS x RA	3.96	0.02	15.23	0.00
TL x SS x DT	0.46	0.77	1.87	0.11
TL x SS x RA	0.29	0.89	0.60	0.67
TL x DT x RA	1.51	0.22	0.73	0.48
SS x DT x RA	0.71	0.49	1.39	0.25
TL x SS x DT x RA	0.20	0.94	0.24	0.92

Note. ANOVA = analysis of variance.

discrimination (Table 2). There was also an interaction of sample size with relative ability. An analysis of this interaction, shown in Table 12, revealed that, in the presence of unequal ability, errors significantly increased with sample size. With equal ability, they remained constant across sample size.

Table 12

Interaction Cell Means for Logistic Regression Method and False Positive Errors

	Relative ability		
Sample size	Equal	Unequal	*F*
200	6.25	6.72	2.91
400	6.13	7.61	28.60*
600	6.41	9.04	90.36*

*$p < .001$.

Summary

In the detection of uniform, non-uniform, and combination DIF, the detection rates of both studied methods were not affected by varying test length. Both methods detected all DIF types at higher rates with larger sample sizes.

When the ANOVA DIF method was used, the experimental factor discrimination type interacted with sample size in the detection of uniform DIF, and the detection rate was higher with equal underlying ability than with unequal underlying ability. The ANOVA DIF method applied to non-uniform DIF resulted in a low detection rate overall and revealed an interaction between discrimination type and relative ability. In the detection of combination DIF, a three-way interaction among sample size, discrimination type, and relative ability was found. Errors were more likely with 20-item tests, larger sample sizes, and varying discrimination.

When the logistic regression method was used to detect uniform DIF, the detection rate was higher with equal rather than unequal underlying ability between groups. With uniform and non-uniform DIF, the logistic regression method had a higher detection rate with larger sample sizes as well as with items with varying discrimination. Like the ANOVA DIF method, the logistic regression method applied to combination DIF resulted in a three-way interaction among sample size, relative ability, and discrimination type. The error rate was higher with varying discrimination than with fixed and increased with sample size in the presence of unequal relative underlying ability.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

Conclusions

For the logistic regression and analysis of variance (ANOVA) differential item functioning (DIF) methods, the research questions investigated the presence of significant interactions or main effects for the experimental factors in the detection of uniform, non-uniform, and combination DIF, as well as false positive errors. in the case of uniform DIF, the ANOVA method had a significant interaction between sample size and discrimination type, with relative ability a main effect. In the detection of uniform DIF with logistic regression, there were no interactions between any of the factors, and sample size, discrimination type, and relative ability were main effects. For non-uniform DIF, the ANOVA method had a significant interaction between discrimination type and relative ability, and sample size was a main effect. In the detection of non-uniform DIF with logistic regression, there were no interactions among any of the factors, and sample size and discrimination type were significant main effects. For combination DIF, there was a significant three-way interaction among sample size, discrimination type, and relative ability for both methods, and there were

no main effects. For false positive errors, the ANOVA method had no interactions among factors, and test length, sample size, and discrimination type were main effects. The logistic regression method had a significant interaction between sample size and relative ability, and discrimination type was a main effect.

Two general conclusions encompass the detection of uniform DIF, non-uniform DIF, and the combination of these two types of DIF, regardless of the detection method. First, the detection rate improved significantly as sample size increased. This is in keeping with Tang's (1994) findings regarding the ANOVA DIF method in detecting uniform DIF, Rogers and Swaminathan's (1993) findings regarding logistic regression in detecting uniform, non-uniform, and combination DIF, and Tian, Pang, and Boss's (1994) findings regarding logistic regression in DIF detection with empirical data. Second, the detection rate did not change significantly between tests of different lengths. This is in accordance with Rogers and Swaminathan (1993), who found that, with logistic regression and test lengths of 40 and 80, test length had no effect on uniform DIF detection rates and only a slight effect (a 5% difference) on non-uniform DIF detection rates. This result was also expected for the ANOVA DIF method because a test length of 20 items is adequate for applications based on the Rasch measurement model (Wright & Stone, 1979). Otherwise, results varied between types of DIF.

Uniform Differential Item Functioning

With the ANOVA method, the ordinal interaction between sample size and discrimination type was due to a decrease in the detection rate difference between discrimination types at sample size 600. Thus, with a sufficiently large sample, the difference in detection rates between discrimination types may disappear. Relative ability was found to be a main effect, with a higher detection rate for equal ability than for unequal ability. With logistic regression, there were no interactions among any of the factors, and sample size, discrimination type, and relative ability were main effects. The detection rate increased with sample size, was higher for varying discrimination than for fixed, and was higher for equal ability than for unequal ability.

In using either detection method, equal ability, larger sample sizes, and varying discrimination increased the detection rate. The increased detection rate in the presence of varying discrimination should be viewed with caution, however. In this particular study, the discrimination parameter values for items with varying discrimination tended to be larger than the discrimination parameter values for items with fixed discrimination. In the presence of uniform DIF, items with high discrimination values will have more area between the item characteristic curves of comparison groups than will items with low discrimination. Therefore, the procedures could have been taking advantage of more DIF rather than of varying discrimination. However, this seems unlikely, as

the increased detection rate for varying discrimination was much less pronounced or nonexistent with non-uniform or combination DIF.

Non-Uniform Differential Item Functioning

The ANOVA DIF method's interaction between discrimination type and relative ability indicated that unequal underlying ability, combined with varying discrimination (a form of misfit to the Rasch model), will significantly reduce the detection rate. Schulz, Perlman, Rice, and Wright (in press) noted that, when underlying abilities between groups are unequal, Rasch-based methods of DIF detection may not be able to separate non-uniform DIF from other sources of misfit to the Rasch model. This certainly seems true in the case of the ANOVA DIF method. In contrast, there were no interactions with the logistic regression method, and the only significant main effect (besides the universal one--sample size) was a slight increase in DIF detection when discrimination varied.

The detection of non-uniform DIF was expected to exhibit the greatest differences between the detection procedures because the logistic regression technique was designed to detect non-uniform DIF and the ANOVA procedure was not. The overall ANOVA DIF detection rate was half that of the logistic regression method.

Combination Differential Item Functioning

A significant three-way interaction among sample size, discrimination type, and relative ability was present for both DIF detection methods. No main effects were found with the use of either method. The methods seemed to behave most similarly in the detection of combination DIF, as opposed to the detection of exclusively uniform or exclusively non-uniform DIF.

The higher detection rate when underlying ability is unequal rather than equal between comparison groups was a surprise. When DIF was exclusively uniform, it was detected at a higher rate with equal relative ability than with unequal relative ability. With non-uniform DIF, the detection rates differed only slightly between types of relative ability. It is possible that the broader range of ability level used when underlying ability was unequal gave better information to the procedures than when underlying ability was equal; however, this does not explain whey the results for combination DIF are unlike those for exclusively uniform and exclusively non-uniform DIF. Perhaps this result occurred only with combination DIF because the area between the reference and focal groups' item characteristic curves was larger than with only one DIF type present. This result could be attributed in part to the presence of more DIF rather than the simultaneous presence of two DIF types.

False Positive Errors

With the ANOVA DIF method, no significant interactions existed among any of the experimental factors, and test length, sample size, and discrimination type were all main effects. With logistic regression, there was a significant interaction between sample size and relative ability, and discrimination type was a main effect.

Tang (1994) found an increase in errors with the ANOVA DIF method when underlying ability was unequal between comparison groups as sample size increased. In contrast, this study found relative underlying ability to be the only insignificant factor studied with respect to the ANOVA DIF method and false positive errors. Tests of lengths of 40 and 60 had a smaller error rate than those with 20 items, whereas the rate increased with increased sample size. Both methods had higher error rates in the presence of varying as opposed to constant discrimination. Curiously, the logistic regression method exhibited the same interaction between sample size and relative ability that Tang (1994) found when using the ANOVA DIF method.

Recommendations

Suggestions for Differential Item Functioning Detection

The lack of any interaction between test length and sample size indicates that either method can be utilized with all combinations of test length and sample size contained herein, although both methods had

higher detection rates with larger sample sizes. For both methods, a test length of 20 was adequate. The performance of either method with shorter tests is unknown, although 20 is the minimum test length recommended for Rasch model applications such as the ANOVA DIF method. Whereas a few specific combinations of experimental factors and DIF types had adequate results with a sample size of 400, in almost all situations 600 was significantly better. Larger sample sizes than those contained in this study have been examined for both detection methods in the literature. The smallest sample size in this study (200 persons) is not recommended due to low detection rates with either method. Accordingly, sample sizes below 200 are also proscribed. Thus, the search for a small-sample DIF detection method outside the realm of classical true-score measurement theory does not seem to have ended.

Looking beyond the two factors, test length and sample size, the effects of discrimination type and relative ability depend on what type of DIF is to be detected. And, although a practitioner can easily discover the underlying ability of comparison groups as approximated by test scores, the determination of discrimination type is not an easily solved problem. Thus, in practice, the knowledge of discrimination type may not be available. With combination DIF and non-uniform DIF, the logistic regression method was less affected than the ANOVA DIF method by differences in discrimination type, especially as sample size increased.

Any given test could harbor a variety of combinations of levels of the experimental factors examined in this study. The logistic regression method had fewer interactions among the experimental factors, and it would yield more interpretable and accurate results if a practitioner was uncertain about the nature of the test items to be studied. The detection rates were also generally higher than for the ANOVA DIF method.

In the detection of uniform DIF, a test consisting of items with varying discrimination could be safely analyzed with the ANOVA DIF method, in spite of the model-data misfit that would exist. Unfortunately, to assure that only non-uniform DIF was present, items that did not fit the Rasch model would need to be eliminated from the test (Smith, 1993), which would obviate the usefulness of the method's apparent robustness to model-data misfit.

Practitioners who have computed item statistics for a test using a Rasch model would have only a few additional calculations to perform in order to apply the ANOVA DIF detection method. In contrast, the logistic regression method does not build upon other item analysis procedures.

If the items are going to be subjected to a Rasch analysis to obtain item statistics, and if the item were found to fit the Rasch model, then the ANOVA DIF method is recommended. However, if there is a chance that non-uniform DIF is present in the data, and if it is important that it be detected, the ANOVA DIF method should be avoided.

When underlying ability between comparison groups is not equal, this study found that the logistic regression method's false positive error rate increased with sample size. Because Tang (1994) found the same result with the ANOVA DIF method, caution with either method is advised in the presence of unequal ability.

The ANOVA DIF method had not been studied in the context of any DIF type besides uniform or any discrimination type besides varying. The interactions among discrimination type and other experimental factors in the detection of all three DIF types indicates that the method may be of little value with items of varying discrimination when a clear analysis of DIF is required. Further, the advantages of a higher detection rate for both varying discrimination and equal relative ability with uniform DIF are reversed with combination DIF. It is highly unlikely that a practitioner will have advance knowledge of the type of DIF present when using any DIF detection method. Thus, the dependence of detection rates on the type of DIF, combined with discrimination type and underlying ability, proscribes the ANOVA DIF detection method in many situations.

The logistic regression method had been shown in prior studies to be acceptable for DIF detection in a variety of situations. The present study adds to the knowledge about this technique's detection rate under previously unstudied conditions. In the case of strictly uniform and strictly non-uniform DIF, the detection rate of 35% to 36% with a

sample size of 200 indicates that the logistic regression technique is of little practical use in extremely small samples. The present study also found that a test length of 20 was adequate for DIF detection with the logistic regression method. There is no clear recommendation with regard to relative ability and discrimination type. For uniform and non-uniform DIF detection, the detection rate was clearly higher with varying discrimination. For uniform DIF, equal underlying ability was associated with a higher detection rate, whereas non-uniform DIF was detected equally well regardless of relative underlying ability. For combination DIF, the results are less clear, due to the interaction among sample size, discrimination type, and relative ability. Tian et al. (1994) presented a modification of the logistic regression procedure, which seems more likely to detect uniform and non-uniform DIF, as well as to identify which type of DIF is present. This procedure is highly recommended.

Suggestions for Future Research

The ANOVA DIF detection method should be investigated in several areas that were beyond the scope of this study:

1. It would be of great interest to see how closely the ANOVA DIF method correlates with other, more established, Rasch-based detection methods, especially in the presence of small sample sizes. Theoretically, the ANOVA DIF method's use of a single calibration should enable it to

work satisfactorily with sample sizes as small as half of those required for other Rasch-based detection methods.

2. It may be possible to use the ANOVA DIF method, even when non-uniform DIF is present, by performing a simple screening technique before DIF detection analysis. Hambleton and Rogers (1989) suggested as two possibilities the comparison of the direction of item difficulties between score groups at different score levels and graphing techniques.

3. The reliability of the ANOVA DIF method has not been assessed using empirical data. The comparison of DIF detection results among samples drawn from a common population could help determine whether or not this method yields consistent results with real data.

APPENDIX A

DATA GENERATION PROGRAM

```
'   MAKEDIF
'   A program to generate scored dichotomous item responses with options
'   for difficulty and discrimination DIF.
'   Last revision October 16, 1994
'   The input file infile$ must contain the following information
'
'     fname, a filename of 8 or fewer characters enclosed in double
'     quotation marks, along with any drive and/or path specifications.
'     This is where output will be written.
'
'     iall, the length of the test.  This number must be 20, 40, or 60.
'
'     grupsize, the size of the focal group.  This number is half of the
'     total group size for which data will be generated.  (This program
'     assumes that reference and focal group sizes are equal.)
'
'     firstrep, the identification number for the first replication.
'
'     lastrep, the identification number for the last replication.
'
'     discfixd, which should equal 1 if item discrimination is fixed
'     and otherwise be 0.
'
'     abilequl, which should equal 1 if reference and focal groups have
'     equal ability and otherwise be 0.
'
'     For example, if the input file consists of the following line:
'     "b:\data\exper23"  40  200  425  430  435    0    1
'     running the program would result in the files exper23.430,
'     exper23.431, exper23.432, exper23.433, exper23.434, and
'     exper23.435 being written to the directory data on drive b.  Each
'     file would be for a 40 item test taken by 200 focal group and 200
'     reference group members, with varying discrimination and equal
'     underlying ability.
'
DECLARE FUNCTION gasdev! ()                          'returns random number from
                                                     'unit normal distribution
DECLARE FUNCTION normdist! (mean!, stdev!)           'returns normally distributed
                                                     'random number with specified
                                                     'mean and standard deviation
DECLARE FUNCTION lognorm! ()
GOSUB getinput
GOSUB verify
'
'                              declare arrays
'
'     alpha holds the discrimination parameter for each item
'     beta holds the difficulty parameter for each item
'     probcorr is the probability an item is answered correctly
'     response is 0 for correct item, 1 otherwise
'     theta holds the ability parameter for each person
'     discdif equals 1 if item discrimination is altered to simulate dif, 0
'             otherwise
'     diffdif equals 1 if item difficulty is altered to simulate dif, 0
'             otherwise
DIM alpha(iall), beta(iall), probcorr(iall), response(iall), theta(pall)
DIM discdif(iall), diffdif(iall)
RANDOMIZE TIMER
GOSUB initarrays
GOSUB assigndif
FOR thisrep = firstrep TO lastrep
  Exten$ = STR$(thisrep)
  toobig = LEN(Exten$)
  Exten$ = RIGHT$(Exten$, toobig - 1)
```

```
  outfile$ = basename$ + "." + Exten$
  PRINT "replication number "; thisrep; " printing to "; outfile$
  OPEN outfile$ FOR OUTPUT AS #2
  GOSUB itemparams
  FOR p = 1 TO pall
    GOSUB genabil
    GOSUB writeabil
    score = 0
    FOR i = 1 TO iall
      GOSUB unifdif
      GOSUB nonunifdif
      GOSUB genscors
      PRINT #2, USING "#"; response(i);
    NEXT i
    PRINT #2, USING "###"; score
  NEXT p
  CLOSE #2
NEXT thisrep
END
getinput:
  INPUT "enter input file name:", infile$: OPEN infile$ FOR INPUT AS #3
  INPUT #3, basename$, iall
  IF iall <> 20 AND iall <> 40 AND iall <> 60 THEN PRINT "#items bad": CLOSE :
STOP
  INPUT #3, grupsize, firstrep, lastrep, discfixd, abilequl
  IF firstrep > lastrep OR lastrep > 999 OR (lastrep - firstrep) > 999 THEN
PRINT "replication numbers bad": CLOSE : STOP
  IF discfixd <> 0 AND discfixd <> 1 THEN PRINT "discfixd flag bad": CLOSE :
STOP
  IF abilequl <> 0 AND abilequl <> 1 THEN PRINT "abilequl flag bad": CLOSE :
STOP
  pall = grupsize * 2
RETURN
verify:
  PRINT basename$; ".";
  PRINT USING "###"; firstrep;
  PRINT " will be initial output file"
  PRINT basename$; ".";
  PRINT USING "###"; lastrep;
  PRINT " will be final output file"
  PRINT iall; " items from 2 groups of "; grupsize; " will be replicated."
  IF discfixd = 1 THEN PRINT "discrimination will be fixed."
  IF discfixd = 0 THEN PRINT "discrimination will vary."
  IF abilequl = 1 THEN PRINT "ability between groups will be equal."
  IF abilequl = 0 THEN PRINT "ability between groups will be unequal."
  INPUT "okay to continue?  y or n", signal$
  IF signal$ <> "y" AND signal$ <> "y" THEN CLOSE : STOP
RETURN
initarrays:
  FOR i = 1 TO iall
    discdif(i) = 0
    diffdif(i) = 0
  NEXT i
RETURN
assigndif:
  diffdif(6) = 1
  discdif(13) = 1
  diffdif(20) = 1: discdif(20) = 1
  IF iall > 20 THEN
    diffdif(26) = 1
    discdif(33) = 1
    diffdif(40) = 1: discdif(40) = 1
  END IF
  IF iall > 40 THEN
```

```
    diffdif(46) = 1
    discdif(53) = 1
    diffdif(60) = 1: discdif(60) = 1
  END IF
RETURN
itemparams:
  FOR i = 1 TO iall
    beta(i) = normdist(0, 1)
    IF discfixd THEN alpha(i) = 1 ELSE alpha(i) = lognorm
  NEXT i
RETURN
genabil:
  IF (abilequl) OR (p > grupsize) THEN theta(p) = normdist(0, 1) ELSE theta(p) =
normdist(-1, 1)
RETURN
writeabil:
  PRINT #2, USING "###"; p;
  PRINT #2, USING "##.### "; theta(p);
RETURN
nonunifdif:
SELECT CASE discdif(i)
CASE 1
  IF (p <= grupsize) THEN alphahat = 1.763073 ELSE alphahat = 1
CASE 0
  alphahat = alpha(i)
CASE ELSE
  PRINT "Program array discdif improperly filled.  Fatal error.": CLOSE : STOP
END SELECT
RETURN
unifdif:
  IF p <= grupsize AND diffdif(i) THEN betahat = beta(i) + .6 ELSE betahat =
beta(i)
RETURN
genscors:
  probcorr(i) = 1 / (1 + EXP(-1 * alphahat * (theta(p) - betahat)))
  IF RND <= probcorr(i) THEN
    response(i) = 1
    score = score + 1
  ELSE response(i) = 0
  END IF
RETURN
FUNCTION gasdev
'
'From Sprott, 1991
'
STATIC iset, gset
SELECT CASE iset
CASE 0
  DO
    v1 = 2! * RND - 1!
    v2 = 2! * RND - 1!
    r = v1 ^ 2 + v2 ^ 2
  LOOP WHILE r >= 1! OR r = 0!
  fac = SQR(-2! * LOG(r) / r)
  gset = v1 * fac
  gasdev = v2 * fac
  iset = 1
CASE ELSE
  gasdev = gset
  iset = 0
END SELECT
END FUNCTION
FUNCTION lognorm
  z = normdist(-.065, .13)
```

```
  lognorm = 1.7 * EXP(z)
END FUNCTION
FUNCTION normdist (mean, stdev)
  normdist = mean + stdev * gasdev
END FUNCTION
```

APPENDIX B

RASCH RESIDUAL SCORE CALCULATION PROGRAM

```
'       RASCHRES:  A program to estimate Rasch residual scores.
'
'       This program reads in dichotomously-scored responses, then
'       computes expected and residual scores using the Rasch model and
'       UCON estimation.  Residual scores match BIGSTEPS output to 2
'       decimal places.
'
'       From J. M. Linacre."Designing your own Rasch analysis program"
'                    ERIC document ED 318 801, April, 1990.
'       Also F. B. Baker, 1992.
'       Last revision, 16-NOV-94.
'
'       datafile$                input file
'       outfile$                 output file
'       iall                     number of items in datafile$
'       pall                     number of persons in datafile$
'       I                        index variable of item being calibrated
'       p                        index variable of person being measured
'       iexp(i), pexp(p)         holds expected item and person scores
'       ilogit(i), plogit(p)     holds logit calibration or measure
'       iscore(i), pscore(p)     holds number of successes
'       ivar(i), pvar(p)         holds variance of logit estimate
'
'       Command file format:     Line 1  iall, pall, startcol
'                                Line 2  "INPUT.FIL"      "OUTPUT.FIL"
'                                Line 3  "MOREINP.UT"    "MOREOUT.PUT"
'                                etc.
'                                The number of lines is limited only by
'                           the number of input files and available disk
'                                space for output files.  All input files
'                                contain IALL items and PALL people,
'                           and the data should begin in column STARTCOL
'
'  Input file format:       Scored responses, no spaces between columns,
'                           beginning in column STARTCOL.  One line per
'                                person.  No blank lines.
'                                1 = correct; 0 = incorrect.
'
'  Output file format:      Each person's first line is the information
'                                read from input file.  Subsequent lines
'                                consist of residual scores computed to 3
'                          places, with a maximum line length of 120 cols.
'
CLS
INPUT "enter command file name:", commfile$:
OPEN commfile$ FOR INPUT AS #3
INPUT #3, iall, pall, startcol
time1 = TIMER
DIM icount(iall), iexp(iall), ilogit(iall), iscore(iall), ivar(iall)
DIM pcount(pall), pexp(pall), plogit(pall), pscore(pall), pvar(pall)
DIM response(pall, iall)
DIM raschres(pall, iall)
oneless = startcol - 1
DO UNTIL EOF(3)
  itertime = TIMER
'       initialize arrays
'
```

```
  FOR i = 1 TO iall: icount(i) = 0: iscore(i) = 0: ilogit(i) = 0: NEXT i
  FOR p = 1 TO pall: pcount(p) = 0: pscore(p) = 0: plogit(p) = 0: NEXT p
'
  FOR p = 1 TO pall
    FOR i = 1 TO iall
      raschres(p, i) = 9
    NEXT i
  NEXT p
'
itotal = iall: ptotal = pall
'
'       read data file, beginning in column startcol
'
  INPUT #3, datafile$, outfile$
  OPEN datafile$ FOR INPUT AS #1
  OPEN outfile$ FOR OUTPUT AS #2
  LOCATE 1, 1
  PRINT "datafile = "; datafile$; "  outfile = "; outfile$
  FOR p = 1 TO pall: LINE INPUT #1, l$
    FOR i = 1 TO iall
      r$ = MID$(l$, oneless + i, 1)
      IF r$ <> "0" AND r$ <> "1" THEN
        response(p, i) = -1
      ELSE
        response(p, i) = VAL(r$)
      END IF
    NEXT i
  NEXT p
'
recount = -1
'     This comment replaces code which was used when this program was
'     actually run.  To make this program operational, insert the last
'     13 lines from Linacre (1990) page 17 (ERIC page 18) and the first
'     19 lines from page 18 (ERIC page 19) in place of this comment.
WEND

GOSUB proxest
'
'       PROX algorithm first.
'       Estimates converge when no measure changes by more than 0.1
logits.
'
'
converged$ = "no"
est = 0
LOCATE 23, 1
PRINT "
 "
WHILE converged$ = "no": converged$ = "yes"
  LOCATE 23, 1
  est = est + 1: PRINT "est # "; est
  FOR i = 1 TO iall: iexp(i) = 0: ivar(i) = 0: NEXT i
'     This comment replaces code which was used when this program was
'     actually run.  To make this program operational, insert the last
'     43 lines from Linacre (1990) page 18 (ERIC page 19) in place of
'     this comment.
WEND: PRINT "completed"
```

```
'
ibias = (itotal - 1!) / itotal
FOR i = 1 TO iall
  IF icount(i) > 0 THEN ilogit(i) = ilogit(i) * ibias
NEXT i
'
pbias = (ptotal - 1!) / ptotal
FOR p = 1 TO pall
  IF pcount(p) > 0 THEN plogit(p) = plogit(p) * pbias
NEXT p
'
FOR p = 1 TO pall
  IF pcount(p) > 0 THEN
    FOR i = 1 TO iall
      IF icount(i) > 0 AND response(p, i) >= 0 THEN
        success = 1! / (1! + EXP(ilogit(i) - plogit(p)))
        variance = success * (1! - success)
        raschres(p, i) = response(p, i) - success
      END IF
    NEXT i
  END IF
NEXT p
'
'       write results to disk
'
LOCATE 23, 40
PRINT "                          "
FOR p = 1 TO pall
  LOCATE 23, 40
  PRINT "saving person "; p
  PRINT #2, USING "###"; pscore(p);
  FOR i = 1 TO iall
    PRINT #2, USING "##.###"; raschres(p, i);
    IF (i MOD 20) = 0 THEN PRINT #2,                  'limit output to
120 cols
  NEXT i
NEXT p
time2 = TIMER
iterlaps = time2 - itertime
elapsed = time2 - time1
LOCATE 24, 1
PRINT " time this iteration "; iterlaps; "total "; elapsed
CLOSE #1
CLOSE #2
LOOP
CLOSE #3
END

proxest:
'This algorithm is from Baker, 1992
proxsum = 0
FOR i = 1 TO iall
  IF icount(i) > 0 THEN
    partone = ptotal / iscore(i)
    ilogit(i) = LOG(partone - 1)
    proxsum = proxsum + ilogit(i)
  END IF
```

```
NEXT i
meanl = proxsum / itotal
FOR i = 1 TO iall
  IF icount(i) > 0 THEN ilogit(i) = ilogit(i) - meanl
NEXT i
FOR p = 1 TO pall
  IF pcount(p) > 0 THEN
    parttwo = itotal - pscore(p)
    plogit(p) = LOG(pscore(p)) - LOG(parttwo)
  END IF
NEXT p
RETURN
```

APPENDIX C

RESIDUAL SCORE ANALYSIS OF VARIANCE PROGRAM

```
DECLARE FUNCTION betai! (a!, b!, x!)
DECLARE FUNCTION betacf! (a!, b!, x!)
DECLARE FUNCTION GAMMLN! (x!)
'Last Revised 11-22-94
' input contains iall pall "outfile" and a list of input files.
DECLARE SUB summer (datq!(), n!, sum!, sumsq!, count!)
DECLARE SUB ftest (data1!(), n1!, data2!(), n2!, f!, dfbet!, dfwith!,
probf!)
DECLARE SUB hitcount (indx, u, nu, comb, fp)
CLS
DATA 76.18009173D0,-86.50532033D0,24.01409822D0
DATA -1.231739516D0,.120858003D-2,-.536382D-5,2.50662827465D0
DATA 0.5D0,1.0D0,5.5D0
'DATA lines are used in function GAMMLN
INPUT "Enter command filename:"; commfile$: OPEN commfile$ FOR INPUT AS
#3
INPUT #3, iall, pall, outfile$
IF (iall <> 20) AND (iall <> 40) AND (iall <> 60) THEN
  PRINT "Program requires 20,40, or 60 items, not "; iall
  CLOSE
  END
END IF
grupsize = pall / 2
DIM res(pall, 20), grup1(grupsize), grup2(grupsize)
time1 = TIMER
fphitsum = 0
uhitsum = 0
nuhitsum = 0
cbhitsum = 0
OPEN outfile$ FOR OUTPUT AS #2
DO UNTIL EOF(3)
  itertime = TIMER
  GOSUB initarrays
  icount = 0
  fphit = 0
  uhit = 0
  nuhit = 0
  cbhit = 0
  INPUT #3, infile$
  PRINT "Input file = "; infile$, "Output file ="; outfile$
  OPEN infile$ FOR INPUT AS #1
  FOR p = 1 TO pall
    GOSUB GetLine1
  NEXT p
  CLOSE #1
  GOSUB ProcessLine
  IF iall > 20 THEN
    OPEN infile$ FOR INPUT AS #1
    FOR p = 1 TO pall
      GOSUB GetLine2
    NEXT p
    CLOSE #1
    GOSUB ProcessLine
  END IF
  IF iall > 40 THEN
    OPEN infile$ FOR INPUT AS #1
    FOR p = 1 TO pall
```

```
      GOSUB GetLine3
    NEXT p
    CLOSE #1
    GOSUB ProcessLine
  END IF
LOCATE 1, 1
  PRINT #2, "Data taken from "; infile$; " alpha = .05 with "; dfw; "
df."
  PRINT #2, "Type I", "Uniform", "Nonuniform", "Combination"
  PRINT #2, fphit, uhit, nuhit, cbhit
  PRINT #2,
  fphitsum = fphitsum + fphit
  uhitsum = uhitsum + uhit
  nuhitsum = nuhitsum + nuhit
  cbhitsum = cbhitsum + cbhit
  GOSUB ShowTime
LOOP
PRINT #2,
PRINT #2, "                                    TOTALS"
PRINT #2, " ", "Type I", "Uniform ", "Nonuniform", " Combination"
PRINT #2, "Found", fphitsum, uhitsum, nuhitsum, cbhitsum
fpdenom = 85 * iall
difdenom = 5 * iall
PRINT #2, "Possible", fpdenom, difdenom, difdenom, difdenom
PRINT #2, "Percent",
PRINT #2, USING "##.##"; 100 * fphitsum / fpdenom; : PRINT #2, "%",
PRINT #2, USING "##.##"; 100 * uhitsum / difdenom; : PRINT #2, "%",
PRINT #2, USING "##.##"; 100 * nuhitsum / difdenom; : PRINT #2, "%",
PRINT #2, USING "##.##"; 100 * cbhitsum / difdenom; : PRINT #2, "%"
CLOSE #2
CLOSE #3
END
initarrays:
  FOR p = 1 TO pall
    FOR i = 1 TO 20
      res(p, i) = 0
    NEXT i
  NEXT p
  FOR p = 1 TO grupsize
  grup1(p) = 0
  grup2(p) = 0
  NEXT p
RETURN

GetLine1:
  LINE INPUT #1, line$
  startpos = 4
  GOSUB FilMatrx
  IF iall >= 40 THEN LINE INPUT #1, dum$
  IF iall = 60 THEN LINE INPUT #1, dum$
RETURN

GetLine2:
  LINE INPUT #1, dum$
  LINE INPUT #1, line$
  startpos = 1
  GOSUB FilMatrx
```

```
  IF iall = 60 THEN LINE INPUT #1, dum$
RETURN

GetLine3:
  LINE INPUT #1, dum$
  LINE INPUT #1, dum$
  LINE INPUT #1, line$
  startpos = 1
  GOSUB FilMatrx
RETURN

ProcessLine:
  FOR i = 1 TO 20
    fifth = i / 5
    FOR p = 1 TO grupsize
      grup1(p) = res(p, i)
    NEXT p
    FOR p = grupsize + 1 TO pall
      grup2(p - grupsize) = res(p, i)
    NEXT p
    icount = icount + 1
    CALL ftest(grup1(), grupsize, grup2(), grupsize, f, dfb, dfw, probf)
    PRINT #2, USING "##"; icount;
    PRINT #2, USING "###.##"; f;
    IF probf < .05 THEN
      CALL hitcount(i, uhit, nuhit, cbhit, fphit)
    ELSEIF f = -9 THEN
      PRINT #2, "missing";
    ELSE
      PRINT #2, " nosig ";
    END IF
    IF fifth = INT(fifth) THEN PRINT #2,
    NEXT i
RETURN

FilMatrx:
  FOR i = 1 TO 20
    res(p, i) = VAL(MID$(line$, startpos, 6))
    startpos = startpos + 6
  NEXT i
RETURN

ShowTime:
  time2 = TIMER
  iterlaps = time2 - itertime
  elapsed = time2 - time1
  LOCATE 24, 1
  PRINT "time this iteration "; iterlaps, " total "; elapsed; ""
RETURN

FUNCTION betacf (a, b, x)
'
'This subroutine for a complete beta function is available in Sprott
(1991).
'It is called from betai.
'
END FUNCTION
```

```
FUNCTION betai (a, b, x)
'
'This subroutine for an incomplete beta function is available in Sprott
(1991).
'It is used to compute the significance of F values in sub ftest.
'It calls the functions betacf and GAMMLN.
'
END FUNCTION

SUB ftest (data1!(), n1!, data2!(), n2!, f!, dfbet!, dfwith!, probf!)
CALL summer(data1!(), n1!, sum1!, sumsq1!, count1!)
CALL summer(data2!(), n2!, sum2!, sumsq2!, count2!)
ntot = count1 + count2
IF count1 <> 0 AND count2 <> 0 THEN
  sumtot = sum1 + sum2
  sumsqtot = sumsq1 + sumsq2
  ssav1 = sum1 * sum1 / count1
  ssav2 = sum2 * sum2 / count2
  ssavtot = sumtot * sumtot / ntot
  ssbet = ssav1 + ssav2 - ssavtot
  sswith = sumsqtot - ssav1 - ssav2
  dfbet = 1
  dfwith = ntot - 2
  f = ssbet * dfwith / sswith
  probf = betai(.5 * ntot, .5, ntot / (ntot + f))  'algorithm from
Sprott, 1991
  IF probf > 1! THEN probf = 2! - probf
ELSE
  f = -9
  probf = 999
  dfwith = 0
END IF
END SUB
FUNCTION GAMMLN (XX)
'
'This function is available from Sprott, 1991.  It is called from
function betai.
'
END FUNCTION
SUB hitcount (indx, u, nu, comb, fp)
  PRINT #2, " *SIG* ";
  SELECT CASE indx
  CASE 6
    u = u + 1
  CASE 13
    nu = nu + 1
  CASE 20
    comb = comb + 1
  CASE ELSE
    fp = fp + 1
  END SELECT
END SUB

SUB summer (datq!(), n!, sum!, sumsq!, c!)
'
'given array datq of length n, returns sum of elements as sum and
squared
```

```
'elements summed as sumsq
'
sum = 0!
sumsq = 0!
c = 0!
FOR J = 1 TO n
   element = datq(J)
   IF element = 9 THEN
     c = c
   ELSE
     c = c + 1
     sum = sum + element
     sumsq = sumsq + (element * element)
   END IF
NEXT J
END SUB
```

APPENDIX D

LOGISTIC REGRESSION PROGRAMS

```
COMMENT This is experiment 1 replication 1
DATA LIST FILE = 'exp01DAT 1 A1'
  /PERSON 1-3 SCORE1 TO SCORE20 11-30 TOTAL 32-33
COMPUTE GROUP = 0
IF (PERSON < 301) GROUP = 1
LOGISTIC REGRESSION SCORE1 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE2 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE3 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE4 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE5 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE6 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE7 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE8 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE9 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE10 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE11 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE12 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE13 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE14 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE15 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE16 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE17 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE18 WITH TOTAL GROUP GROUP BY TOTAL
```

```
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE19 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL
LOGISTIC REGRESSION SCORE20 WITH TOTAL GROUP GROUP BY TOTAL
 /METHOD ENTER TOTAL
 /METHOD ENTER GROUP GROUP BY TOTAL


COMMENT This assumes the results of the above program are
COMMENT contained in the file EX01R001 LISTING E1
COMMENT This is experiment 1 replication 1
SET ERRORS=BOTH MESSAGES=NONE PRINTBACK=NONE RESULTS=LISTING
INPUT PROGRAM
DATA LIST FILE = 'EX01R001 LISTING E1'
 /LINEFLAG 3-4 (A)
DO IF (LINEFLAG = "-2")
REREAD
DATA LIST RECORDS = 2
  /2 KISQ 23-32(3)
  DF 36-38
  SIGNIF 47-51 (4)
END CASE
END IF
END INPUT PROGRAM
SELECT IF DF = 2
COMPUTE CASE = $CASENUM
COMPUTE ITENUM = MOD(CASE, 20)
LIST VAR = CASE KISQ DF SIGNIF ITENUM
SELECT IF (SIGNIF <= .05)
COMPUTE UDIF=0
COMPUTE NUDIF=0
COMPUTE COMBDIF=0
COMPUTE FALSEPOS=0
DO IF ITENUM=0
COMPUTE COMBDIF=1
ELSE IF ITENUM=6
COMPUTE UDIF=1
ELSE IF ITENUM=13
COMPUTE NUDIF=1
ELSE
COMPUTE FALSEPOS = 1
END IF
LIST VAR = ALL
DESCRIPTIVES VAR=UDIF NUDIF COMBDIF FALSEPOS/STAT=SUM
```

APPENDIX E

DATA ANALYSIS PROGRAM

```
filename summstat 'dissresu.lt!';
data allobs;
infile summstat;
input uda nda cda fpa testlen sampsize disc abil repno udlr ndlr cdlr
fplr;
posshits=testlen/20;
possmiss=testlen-(3*posshits);
udapct=(uda/posshits)*100;
ndapct=(nda/posshits)*100;
cdapct=(cda/posshits)*100;
fpapct=(fpa/possmiss)*100;
udlrpct=(udlr/posshits)*100;
ndlrpct=(ndlr/posshits)*100;
cdlrpct=(cdlr/posshits)*100;
fplrpct=(fplr/possmiss)*100;
run;
proc anova data = allobs;
  class testlen sampsize disc abil;
  model udapct = testlen|sampsize|disc|abil;
run;
proc anova data = allobs;
  class testlen sampsize disc abil;
  model udlrpct = testlen|sampsize|disc|abil;
run;
proc anova data = allobs;
  class testlen sampsize disc abil;
  model ndapct = testlen|sampsize|disc|abil;
run;
proc anova data = allobs;
  class testlen sampsize disc abil;
  model ndlrpct = testlen|sampsize|disc|abil;
run;
proc anova data = allobs;
  class testlen sampsize disc abil;
  model cdapct = testlen|sampsize|disc|abil;
run;
proc anova data = allobs;
  class testlen sampsize disc abil;
  model cdlrpct = testlen|sampsize|disc|abil;
run;
proc anova data = allobs;
  class testlen sampsize disc abil;
  model fpapct = testlen|sampsize|disc|abil;
run;
proc anova data = allobs;
  class testlen sampsize disc abil;
  model fplrpct = testlen|sampsize|disc|abil;
run;
```

APPENDIX F

SUMMARY DATA

Table 13

Summary Data for Logistic Regression and Analysis of Variance Differential Item Functioning Detection Methods (100 Replications)

Experiment	# of items possible	Number of items found: ANOVA	Number of items found: LR
1. Uniform DIF	100	30	24
Non-uniform DIF	100	24	39
Combination DIF	100	64	73
False positive	1,700	97	102
2. Uniform DIF	100	67	55
Non-uniform DIF	100	28	51
Combination DIF	100	83	88
False positive	1,700	106	91
3. Uniform DIF	100	78	67
Non-uniform DIF	100	36	70
Combination DIF	100	89	91
False positive	1,700	124	104
4. Uniform DIF	200	69	54
Non-uniform DIF	200	35	60
Combination DIF	200	121	134
False positive	3,400	191	195
5. Uniform DIF	200	128	110
Non-uniform DIF	200	53	109
Combination DIF	200	155	171
False positive	3,400	204	204
6. Uniform DIF	200	165	140
Non-uniform DIF	200	66	155
Combination DIF	200	166	183
False positive	3,400	234	194

(*table continues*)

Experiment	# of items possible	Number of items found: ANOVA	Number of items found: LR
7. Uniform DIF	300	105	94
Non-uniform DIF	300	66	103
Combination DIF	300	177	195
False positive	5,100	281	307
8. Uniform DIF	300	194	162
Non-uniform DIF	300	100	191
Combination DIF	300	222	254
False positive	5,100	311	275
9. Uniform DIF	300	248	215
Non-uniform DIF	300	135	231
Combination DIF	300	261	286
False positive	5,100	312	310
10. Uniform DIF	100	36	25
Non-uniform DIF	100	19	25
Combination DIF	100	88	81
False positive	1,700	104	114
11. Uniform DIF	100	50	45
Non-uniform DIF	100	31	53
Combination DIF	100	93	89
False positive	1,700	124	112
12. Uniform DIF	100	72	58
Non-uniform DIF	100	51	78
Combination DIF	100	99	97
False positive	1,700	143	144
13. Uniform DIF	200	56	44
Non-uniform DIF	200	45	66

(*table continues*)

Experiment	# of items possible	Number of items found: ANOVA	Number of items found: LR
Combination DIF	200	164	150
False positive	3,400	179	208
14. Uniform DIF	200	124	100
Non-uniform DIF	200	70	131
Combination DIF	200	193	189
False positive	3,400	211	260
15. Uniform DIF	200	120	95
Non-uniform DIF	200	107	156
Combination DIF	200	198	195
False positive	3,400	262	270
16. Uniform DIF	300	73	59
Non-uniform DIF	300	69	122
Combination DIF	300	230	210
False positive	5,100	255	270
17. Uniform DIF	300	135	109
Non-uniform DIF	300	97	183
Combination DIF	300	280	277
False positive	5,100	328	328
18. Uniform DIF	300	189	162
Non-uniform DIF	300	136	249
Combination DIF	300	286	293
False positive	5,100	350	355
19. Uniform DIF	100	60	54
Non-uniform DIF	100	19	36
Combination DIF	100	62	78
False positive	1,700	99	111

(*table continues*)

Experiment	# of items possible	Number of items found: ANOVA	Number of items found: LR
20. Uniform DIF	100	92	86
Non-uniform DIF	100	29	63
Combination DIF	100	82	87
False positive	1,700	135	123
21. Uniform DIF	100	90	87
Non-uniform DIF	100	37	72
Combination DIF	100	79	95
False positive	1,700	135	117
22. Uniform DIF	200	122	109
Non-uniform DIF	200	43	77
Combination DIF	200	117	141
False positive	3,400	197	223
23. Uniform DIF	200	184	169
Non-uniform DIF	200	75	121
Combination DIF	200	156	170
False positive	3,400	227	227
24. Uniform DIF	200	192	182
Non-uniform DIF	200	75	168
Combination DIF	200	171	189
False positive	3,400	235	227
25. Uniform DIF	300	189	163
Non-uniform DIF	300	72	108
Combination DIF	300	192	217
False positive	5,100	277	339
26. Uniform DIF	300	257	238
Non-uniform DIF	300	86	184

(*table continues*)

Experiment	# of items possible	Number of items found: ANOVA	Number of items found: LR
Combination DIF	300	240	260
False positive	5,100	332	313
27. Uniform DIF	300	277	270
Non-uniform DIF	300	120	242
Combination DIF	300	250	278
False positive	5,100	361	358
28. Uniform DIF	100	45	38
Non-uniform DIF	100	9	32
Combination DIF	100	47	62
False positive	1,700	100	1128
29. Uniform DIF	100	73	72
Non-uniform DIF	100	26	70
Combination DIF	100	82	96
False positive	1,700	121	154
30. Uniform DIF	100	74	75
Non-uniform DIF	100	33	77
Combination DIF	100	80	96
False positive	1,700	147	176
31. Uniform DIF	200	90	75
Non-uniform DIF	200	35	82
Combination DIF	200	97	145
False positive	3,400	196	250
32. Uniform DIF	200	149	139
Non-uniform DIF	200	47	126
Combination DIF	200	160	188
False positive	3,400	224	266

(*table continues*)

Experiment	# of items possible	Number of items found: ANOVA	Number of items found: LR
33. Uniform DIF	200	165	162
Non-uniform DIF	200	576	165
Combination DIF	200	168	198
False positive	3,400	280	362
34. Uniform DIF	300	144	119
Non-uniform	300	37	109
Combination DIF	300	159	225
False positive	5,100	275	373
35. Uniform DIF	300	221	197
Non-uniform DIF	300	58	195
Combination DIF	300	234	288
False positive	5,100	361	413
36. Uniform DIF	300	239	246
Non-uniform DIF	300	94	247
Combination DIF	300	262	296
False positive	5,100	420	502

Note. ANOVA = analysis of variance; LR = logistic regression; DIF = differential item functioning.

REFERENCES

Angoff, W. H. (1972, September). *A technique for the investigation of cultural differences*. Paper presented at the annual meeting of the American Psychological Association, Honolulu, HI. (ERIC Document Reproduction Service No. ED 069 686)

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96-116). Baltimore: Johns Hopkins University Press.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum.

Baghi, H., & Ferrara, S. (1989, March). *A comparison of IRT, delta plot, and Mantel-Haenszel techniques for detecting differential item functioning across subpopulations in the Maryland Test of Citizenship Skills*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 324 364)

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Bleistein, C. A. (1986). *Application of item response theory to the study of differential item characteristics: A review of the literature*. Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED 268 160)

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 397-413). Hillsdale, NJ: Lawrence Erlbaum.

Camilli, G., & Shepard, L. A. (1987). The inadequacy of ANOVA for detecting test bias. *Journal of Educational Statistics*, *12*(1), 87-99.

Camilli, G., & Smith, J. K. (1990). Comparison of the Mantel-Haenszel test with a randomized and a jackknife test for detecting biased items. *Journal of Educational Statistics, 15*(1), 53-67.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1991, April). *Examination of various influences on the Mantel-Haenszel statistic.* Paper presented at the annual meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 331 876)

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics, 18*(2), 131-154.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 397-413). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (ETS Research Report No. RR-83-9). Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED 230 566)

Engelhard, G., Jr., Anderson, D., & Gabrielson, S. (1989, April). *An empirical comparison of Mantel-Haenszel and Rasch procedures for studying differential item functioning on teacher certification tests*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 307 280)

Faggen, J. (1987). Golden Rule revisited: Introduction. *Educational Measurement Issues and Practice, 6*(2), 5-8.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (pp. 147-200). New York: Macmillan.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education*, *2*(4), 313-334.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology*, *71*(2), 327-333.

Intraprasert, D. (1986). An investigation of the reliability of five methods for detecting test item bias: An empirical study (Doctoral dissertation, University of North Texas). *Dissertation Abstracts International*, *48*, 113A.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detection item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 117-160). Baltimore: Johns Hopkins University Press.

Jensen, A. P. (1973). An examination of culture bias in the Wonderlich Personnel Test. *Intelligence*, *1*, 51-64.

Jensen, A. P. (1974). How biased are culture-loaded tests? *Genetic Psychology Monographs*, *90*, 185-244.

Linacre, J. M. (1990, April). *Designing your own Rasch analysis program.* Paper presented at the Annual Meeting of the American Educational Research Association, Boston. (ERIC Document Reproduction Service No. ED 318 801)

Linacre, J. M., & Wright, B. D. (1987). *Item bias: Mantel-Haenszel and the Rasch model. Memorandum No. 39.* Chicago: MESA Psychometric Laboratory. (ERIC Document Reproduction Service No. ED 281 859)

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*(2), 109-118.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Marascuilo, L. A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on χ^2 statistics. *Journal of Educational Measurement, 18*(4), 229-248.

Masters, J. R. (1988, April). *A study of differences between what is taught and what is tested in Pennsylvania.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA. (ERIC Document Reproduction Service No. ED 295 989)

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1991, April). *The effect of sample size on the functioning of the Mantel-Haenszel statistic.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago. (ERIC Document Reproduction Service No. ED 331 877)

McAllister, P. H. (1993). Testing, DIF, and public policy. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 389-396). Hillsdale, NJ: Lawrence Erlbaum.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*(2), 105-118.

Microsoft Corporation. (1990). *Microsoft QuickBasic.* Redmond, WA: Author.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement, 16*(4), 381-388.

Osterlind, S. J. (1983). *Test item bias.* Sage University paper series on quantitative applications in the social sciences, 07-030. Beverly Hills: Sage Publications.

Plake, B. S. (1981). An ANOVA methodology to identify biased items that take instructional level into account. *Educational Psychological Measurement, 41*, 365-368.

Plake, B. S., & Hoover, H. D. (1979). An analytical method of identifying biased test items. *Journal of Experimental Education, 48*, 153-154.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207.

Raju, N. S., & Normand, J. (1985). The regression bias method: A unified approach for detecting item bias and selection bias. *Educational and Psychological Measurement, 45*, 37-54.

Rogers, H. J. (1989). A logistic regression procedure for detecting item bias. *Dissertation Abstracts International, 50*, 3928A. (University Microfilms No. 90-11,788)

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105-116.

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

Ryan, K. E. (1991). The performance of the Mantel-Haenszel procedure across samples and matching criteria. *Journal of Educational Measurement*, *28*(4), 325-337.

SAS Institute, Inc. (1985). *SAS user's guide: Statistics* (5th ed.). Cary, NC: Author.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, *16*(3), 143-152.

Scheuneman, J. D. (1990, April). *Assessing the utility of item response theory models: Differential item functioning*. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston. (ERIC Document Reproduction Service No. ED 319 779)

Schulz, E. M., Perlman, C., Rice, W. K., & Wright, B. D. (in press). An empirical comparison of Rasch and Mantel-Haenszel procedures for assessing differential item functioning. In G. Englehard, Jr., & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3). Norwood, NJ: Ablex Publishing.

Seong, T. J. (1990, April). *Reconsideration of the ANOVA method of detecting item bias*. Paper presented at the annual meeting of the American Educational Research Association, Boston. (ERIC Document Reproduction Service No. ED 318 791)

Seong, T. J., & Subkoviak, M. J. (1987, April). *A comparative study of recently proposed item bias detection methods*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC. (ERIC Document Reproduction Service No. ED 281 883)

Shepard, L., & Camilli, G. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, *6*(4), 317-375.

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics*, *9*(2), 93-128.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement*, *22*(2), 77-105.

Smith, R. M. (1993, April). *A comparison of the Rasch separate calibration and between fit methods of detecting item bias*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Sprott, J. C. (1991). *Numerical recipes routines and examples in BASIC*. Cambridge, UK: Cambridge University Press.

SPSS, Inc. (1990). *SPSS reference guide*. Chicago: Author.

Stricker, L. J. (1982). Identifying test items that perform differentially in population subgroups: A partial correlation index. *Applied Psychological Measurement*, *6*(3), 261-273.

Stricker, L. J. (1984). The stability of a partial correlation index for identifying items that perform differentially in subgroups. *Educational and Psychological Measurement*, *44*, 831-837.

Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Craig, R. D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. *Journal of Educational Measurement*, *21*(1), 49-58.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370.

Tang, H. (1994, January). *A new IRT-based small sample DIF method*. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 389-396). Hillsdale, NJ: Lawrence Erlbaum.

Tian, F., Pang, X. L., & Boss, M. W. (1994, April). *The effects of sample size and criterion variable on the identification of DIF by the*

Mantel-Haenszel and logistic regression procedures. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Van der Flier, H., Mellenbergh, G. J., Adèr, H. J., & Wijn, M. (1984). An interative item bias detection method. *Journal of Educational Measurement, 21*(2), 131-145.

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wianer (Eds.), *Differential item functioning* (pp. 123-135). Hillsdale, NJ: Lawrence Erlbaum.

Willson, V. L., Nolan, R. F., Reynolds, C. R., & Kamphaus, R. W. (1989). Race and gender effects on item functioning on the Kaufman Assessment Battery for Children. *Journal of School Psychology, 27*(3), 289-296.

Wright, B. D., Mead, R., & Draba, R. (1976). *Detecting and correcting test item bias with a logistic response model* (RM-22). Chicago: University of Chicago, Statistics Laboratory, Department of Education.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*(3), 185-197.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233-251.