University of Alberta

Reducing Type I Error Rates Using an Effect Size Measure with the Logistic Regression Procedure for DIF Detection

by

Michael G. Jodoin ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Education

Department of Educational Psychology

Edmonton, Alberta

Fall, 1999

0-612-47138-1

University of Alberta

Library Release Form

Name of Author: Michael G. Jodoin

Title of Thesis: Reducing Type I Error Rates Using an Effect Size Measure with the Logistic Regression Procedure for DIF Detection

Degree: Master of Education

Year this Degree Granted: 1999

Michael G. Jodoin
#204, 8503 108 Street
Edmonton, Alberta T6E 6J9

August 19, 1999

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Reducing Type I Error Rates Using an Effect Size Measure with the Logistic Regression Procedure for DIF Detection** submitted by **Michael G. Jodoin** in partial fulfillment of the requirements for the degree of **Master of Education**.

Dr. Mark Gierl (Supervisor)

Dr. Jeff Bisanz

Dr. Steve Hunka

Dr. Bruno Zumbo

August 16, 1999

Abstract

The logistic regression procedure for Differential Item Functioning (DIF) detection is a model-based approach designed to identify both uniform and nonuniform DIF (Swaminathan & Rogers, 1990). Past studies have identified an inflated Type I error rate that is problematic because it results in the inefficient use of testing resources and interferes with understanding the underlying psychology of DIF. Zumbo and Thomas (1996) developed an effect size measure for use with the logistic regression DIF procedure and suggested a classification method based on Cohen's (1988) effect size measures. This study develops a new classification method based on those established for the Simultaneous Item Bias Test (Roussos & Stout, 1996b). A simulation study is also conducted to determine if the effect size measure affects the Type I error and power rates for the logistic regression DIF procedure across sample size, ability distribution, and percentage of DIF items included on a test. Results from this study indicate that the inclusion of the effect size measure can substantially reduce Type I error rates when large sample sizes are used although there is a reduction in power that must also be considered.

Acknowledgements

To Jeff, Steve, Todd, and Bruno: Thank you for your helpful comments on the manuscript. Your comments and probing questions have forced me to clarify my thoughts and their expression.

To my colleagues (and friends) Don and Dianne: I'm sure you will see many of the ideas we've discussed (pre-hoc) in the text that follows – those discussions (with their sometimes practically pragmatic topics) have made my time here much more personally and academically rewarding.

To Mark: A quick look at the professional interests on my CV (Item Response Theory, DIF, and Computer Adaptive Testing) hints at your influence. Thank you for your patient mentoring – I believe you are largely responsible for the opportunities I find myself leaving to pursue.

To Dad, Mom, Floyd, and Angie: Your support and understanding is greatly appreciated. You may not have always 'understood' what I was doing but your genuine interest and unconditional love was more helpful than you know.

To Delaine: You are the blessing for which I am most thankful. I can not imagine my life without your strength, your determination, and your love. May God bless everyone with what I find in you.

Michael

Table of Contents

List of Tables

List of Figures

Literature Review

Introduction

Differential Item Functioning (DIF) procedures are currently the dominant psychometric methods for addressing fairness in standardized achievement, aptitude, certification, and licensure testing (for a review of these procedures see Clauser & Mazor, 1998; Millsap & Everson, 1993). These procedures reflect, in large part, a response to the legal and ethical need to ensure that comparable examinees are treated equally. Generally, examinees are split into two groups. The *reference group* consists of majority or advantaged group members and the *focal group* consists of minority or disadvantaged group members. DIF analysis then involves matching members of the reference and focal groups on a measure of ability to ensure comparable examinees are being compared and implementing statistical procedures to identify group differences on individual test items.

These group differences may take two forms that can be visually represented with item response functions. Most DIF procedures are designed to identify uniform (unidirectional) DIF which occurs when an item favors one group over another throughout the ability continuum. Figure 1 shows item response functions for a typical uniform DIF item where the difficulty of the item for the reference and focal group differ. Occasionally, DIF procedures may identify nonuniform (crossing) DIF which occurs when there is an ability by group membership interaction, but generally DIF procedures are not designed to do so. Figure 2 provides an example of nonuniform DIF where item discrimination differs for the reference and focal groups.

Swaminathan and Rogers (1990) applied the Logistic Regression (LR) procedure to DIF detection. This was a response, in part, to the belief that the identification of both uniform and nonuniform DIF was important. The strengths of this procedure are well documented. It is a flexible model-based approach designed specifically to detect uniform and nonuniform DIF with the capability to accommodate continuous and multiple ability estimates. Furthermore, simulation studies have demonstrated comparable power in the detection of uniform and superior power in the detection of nonuniform DIF compared to the Mantel-Haenszel (MH) and Simultaneous Item Bias Test (SIB) procedures (Li & Stout, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). These studies also identified two major weaknesses in the LR DIF procedure: 1) the Type I error or false positive rate was higher than expected, and 2) the lack of an effect size measure.

In the context of DIF, a Type I error is the incorrect identification of an item as displaying DIF when, in fact, it does not. Type I errors are problematic for two reasons. First, the incorrect identification of DIF items could lead to the removal of satisfactory items resulting in the inefficient use of limited testing resources. Second, it could interfere with the development of a better understanding of the nature or underlying psychology associated with DIF. In fact, the Type I error inflation under several commonly occurring situations was severe enough to lead to a third problem: It made meaningful power comparisons between MH, SIB, and LR DIF procedures problematic (Li & Stout, 1996).

Another disadvantage of the LR DIF procedure is the use of a statistical test without an associated effect size measure (i.e., a descriptive statistic for the degree or magnitude of DIF). The use of null hypothesis significance testing in the absence of effect size measures has been, and continues to be, scrutinized (e.g., Cohen, 1988, 1990, 1992, 1994; Kirk, 1996). That is, since the sensitivity or power of the statistical test of a hypothesis is dependent on the sample size employed, a measure to distinguish statistical significance from practical significance or meaningfulness is vital to this type of research. As Potenza and Dorans (1995) noted, "to be used effectively, a DIF detection technique needs an interpretable measure of the amount of DIF" (p.33).

One explanation for the inflated Type I error rate associated with the LR DIF procedure is that statistically significant DIF was being flagged and, by implication, misinterpreted as practically significant DIF. This explanation seems reasonable since the associated chi-square statistic is sensitive to large sample sizes. Thus, large sample sizes may result in high power that is identifying small but non-zero DIF. This seems to be a plausible explanation since several studies have demonstrated high power and increasingly inflated Type I error rates for the LR DIF procedure as the sample sizes of the reference and focal groups became larger (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). In addition, the majority of studies to date have made comparisons of Type I error and power rates for procedures based only on their statistical tests. In the one exception identified, the use of an effect size measure in conjunction with a statistical test reduced Type I errors with the MH and SIB procedures (Roussos & Stout, 1996b). Unfortunately, this study did not consider the potential effects the inclusion of the effect size measure could have on power. Moreover, the LR DIF procedure was not considered as an effect size measure had not yet been proposed. This leaves a void in the DIF literature.

Recently, Zumbo and Thomas (1996; also see Zumbo, 1999) proposed $R^2\Delta$, a weighted least squares effect size measure for the LR DIF procedure, that could be used to quantify the magnitude of uniform or nonuniform DIF in items. Implementing Cohen's (1992) small, medium, and large effect size guidelines, they used $R^2\Delta$ to interpret selected items from an example data set. However, this effect size measure has undergone little additional investigation. At least two factors are worthy of further consideration. First, a systematic comparison between $R^2\Delta$ and existing DIF effect size measure classification guidelines needs to be conducted to either justify the use of the existing or establish new guidelines. Second, an investigation into the ability of $R^2\Delta$ to reduce Type I errors in the LR DIF procedure may provide a preferable alternative to the alpha adjustments suggested in Narayanan and Swaminathan (1996) that would reduce the power of DIF procedures.

Therefore, the purpose of this study is twofold. First, following a review of the LR DIF procedure developed by Swaminathan and Rogers (1990) with a suggested modification to improve the power of uniform DIF detection and the effect size measure proposed by Zumbo and Thomas (1996), an empirical standard setting approach to classify negligible, moderate, and large DIF is conducted. Second, a simulation study of the Type I error and power rates for the chi-square statistical test alone and in conjunction with the proposed modification and effect size measure is presented. Implications for the substantive review of DIF and future simulation studies are also discussed.

The Logistic Regression DIF Procedure

The probability of a correct response to an item using the LR model for the identification of DIF is given by Swaminathan and Rogers (1990):

$$P(u = 1 \mid \theta, g) = \frac{e^{\tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g)}}{1 + e^{\tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g)}}. \quad (1)$$

In this model, θ is the ability or observed trait level of an examinee usually denoted by total test score. Group membership of the examinee, g, is typically coded 1 or 0 for an examinee belonging to the reference or focal group, respectively. The parameters τ_0, τ_1, τ_2, and τ_3 represent the intercept and the weights for the ability, group difference, and ability by group interaction terms, respectively.

Uniform DIF occurs when $\tau_2 \neq 0$ and $\tau_3 = 0$. Furthermore, the uniform DIF favors the reference group when $\tau_2 > 0$ and the focal group when $\tau_2 < 0$. Nonuniform DIF is present when $\tau_3 \neq 0$ regardless of the value of τ_2. When $\tau_3 > 0$, the item favors higher ability members of the reference group and lower ability members of the focal group. In contrast, items with negative values for τ_3 favor higher ability members of the focal group and lower ability members of the reference group.

The null and alternative hypotheses for the simultaneous test of uniform and nonuniform DIF are H_0: $\tau_2 = \tau_3 = 0$ and H_1: $\tau_2 \neq 0$ or $\tau_3 \neq 0$, respectively. The difference between the –2 log likelihood of the compact model (including τ_0 and $\tau_1\theta$ only) and the augmented model $[\tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g)]$ is associated with a chi-square distribution with two degrees of freedom. Typically, when the chi-square test statistic surpasses $\chi^2_{2;\alpha}$, the hypothesis of no DIF is rejected and the item is sent for review by content specialists.

LR was the first procedure purposefully designed to identify both uniform and nonuniform DIF. The two degree of freedom chi-square test was designed to maximize the ability to identify both uniform and nonuniform DIF and control the overall Type I error rate. It was important to control Type I errors given the multiple tests that are conducted in DIF analyses and the inflated false positive rate discussed earlier. Furthermore, there was incomplete information on the prevalence of nonuniform DIF. The intervening years have provided a clearer notion on the nature of DIF. At present, it is commonly acknowledged that nonuniform DIF does occur although with substantially lower frequency than uniform DIF (Camilli & Shepard, 1994; Gierl, Rogers, & Klinger, 1999). Therefore, it seems appropriate to frame DIF tests to focus on uniform DIF but not at the exclusion of nonuniform DIF.

Swaminathan and Rogers (1990) noted, "In the LR model, the interaction term may adversely affect the power of the procedure when only uniform DIF is present because one degree of freedom is lost unnecessarily" (p. 366). Working under the premise that an effective effect size measure can control Type I errors, it seems reasonable to modify the two degree of freedom chi-square test into separate one degree of freedom tests of uniform and nonuniform DIF, respectively. Theoretically, this change should result in superior power in the detection of uniform DIF and nominal Type I errors if an appropriate effect size measure is available. This would enable practitioners to ensure high standards of equity by enhancing the detection of the most common form of DIF in addition to considering nonuniform DIF. However, this will only be prudent if the effect size measure adequately controls Type I errors providing the efficiency required in a testing program.

$R^2\Delta$: A Weighted Least Squares Effect Size Measure for LR

As Zumbo & Thomas (1996) note, the LR model for the identification of DIF given in Equation 1 is nonlinear with respect to the odds or probability. Equivalently, it can be expressed as linear with respect to the odds by

$$\ln\left(\frac{P}{1-P}\right) = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g), \qquad (2)$$

where P is the probability of responding correctly given θ, and g (u=1). This equation can then be considered a weighted least squares model by applying Pregibon's (1981) result that the vector of maximum likelihood estimators $\tilde{\tau}$, of the LR coefficients in equations 1 and 2, can be expressed in terms of weighted least squares by

$$\tilde{\tau} = (X'VX)^{-1}X'Vz \qquad (3)$$

where, $z = X\tilde{\tau} + V^{-1}r$, $r = (u - \tilde{\mathrm{P}})$, V is a NxN diagonal matrix with elements $\tilde{P}_i(1-\tilde{P}_i), i = 1,\ldots,N,$, X is a Nx4 data matrix with rows [1, θ_i, g_i, $\theta_i g_i$], $\tilde{\mathrm{P}}$ is a Nx1 vector of the fitted values of the LR model, u is a Nx1 vector of examinee responses, and N is the combined sample size of the reference and focal groups.

Given the LR DIF procedure could be considered a weighted least squares model, Zumbo and Thomas (1996) extended Pratt's (1987) demonstration that an additive partitioning of the explanatory variables in a LR was reasonable through the geometry of least squares. Furthermore, they applied it specifically to the LR DIF procedure noting that the contribution of each explanatory variable could be defined by

$$R^2\Delta = R^2{}_1 - R^2{}_2, \tag{4}$$

where $R^2{}_1$ and $R^2{}_2$ are the sums of the products of the standardized regression coefficient for each explanatory variable and the correlation between the response and each explanatory variable (i.e., $\sum_{1}^{j} \beta_j r_j$ for j explanatory variables) for the augmented and compact models, respectively. Substantively, the $R^2\Delta$ values corresponding to the uniform and nonuniform terms in the LR DIF procedure, $\tau_2 g$ and $\tau_3(\theta g)$ respectively, could be interpreted as a quantification of magnitude of uniform and nonuniform DIF present in an item and will be referred to as $R^2\Delta$-U and $R^2\Delta$-N [1].

Classification of Negligible, Moderate and Large DIF Using $R^2\Delta$

Based on Cohen's (1992) conventions for small, medium, and large effects, Zumbo and Thomas (1996) suggested a negligible, moderate, and large classification method for $R^2\Delta$. They proposed $R^2\Delta$ values below 0.13 for negligible DIF, between 0.13 and 0.26 for moderate DIF, and above 0.26 for large DIF. Both the moderate and large categories also required the item to be flagged as statistically significant with the two degree of freedom chi-square test.

Both MH and SIB have established effect size measures with criteria to distinguish negligible, moderate, and large DIF that are well accepted (Roussos & Stout

1996b; Zwick & Ercikan, 1989). Although MH may be considered the 'gold standard' in DIF detection (Roussos & Stout, 1996a, 1996b), SIB was chosen as the basis of comparison because it is able to detect both uniform and nonuniform DIF, and has been demonstrated to have superior statistical characteristics in comparison to MH in both uniform and nonuniform DIF detection (Narayanan & Swaminathan, 1994; Roussos & Stout, 1996b; Shealy & Stout, 1993).

To determine if there was a relationship between the SIB effect size measure, $\hat{\beta}_U$, and $R^2\Delta$, four data sets from a large-scale achievement testing program in Canada were examined. These data sets contained a range of DIF effect sizes when the SIB procedure was used. The first two data sets considered English-French translation DIF for 50-item Grade 6 Social Studies and Mathematics achievement tests with 2200 examinees in both the reference and focal groups. Similarly, the second two data sets considered gender DIF for 70-item Grade 12 Social Studies diploma exams. For each data set, $\hat{\beta}_U$ and $R^2\Delta$-U values were calculated. $\hat{\beta}_U$ can be both positive and negative indicating whether the focal or reference group is favored whereas $R^2\Delta$-U is always positive with the direction of advantage indicated by τ_2, as discussed previously. Therefore, the absolute value of $\hat{\beta}_U$ was used in order to facilitate comparison with $R^2\Delta$-U. Bivariate scatterplots of each individual and the combined data sets revealed a consistent curvilinear relationship. Subsequently, cubic curve regression was conducted to predict $R^2\Delta$-U from $\hat{\beta}_U$ with the combined data set because it contained the most data. Furthermore, the cubic model provided the smallest error term of the models fit to the data. Figure 3 portrays the scatterplot and superimposed cubic regression curve for the combined data sets.

Roussos and Stout (1996b) suggested $\hat{\beta}_U$ classification values of 0.059 to distinguish negligible from moderate DIF, and 0.088 to distinguish moderate from large DIF. These $\hat{\beta}_U$ values were substituted into the cubic regression equation to calculate classification scores for $R^2\Delta$-U^2. This procedure indicates the following criteria be used with $R^2\Delta$-U:

- Negligible or A-level DIF: $R^2\Delta\text{-U} < 0.035$,
- Moderate or B-level DIF: Null hypothesis is rejected and $0.035 \leq R^2\Delta\text{-U} < 0.070$,
- Large or C-level DIF: Null hypothesis is rejected and $R^2\Delta\text{-U} \geq 0.070$.

These values are also suggested for use with $R^2\Delta$-N since criteria for the classification of nonuniform DIF effect size measures have not yet been developed by the psychometric community.

The large difference between the classification scores for $R^2\Delta$ suggested by Zumbo and Thomas (1996) and those derived from a comparison with $\hat{\beta}_U$ are striking. In order to investigate the utility of the classification scores developed above, a simulation study was designed to consider the Type I error and power rates of the LR DIF procedure.

Method

Simulation Study

Examinee response data were simulated under a variety of conditions expected to affect the Type I error and power rates of DIF procedures. Three factors were manipulated: sample size, ability distribution differences, and percentage of items containing DIF. Furthermore, the levels of each factor were designed to reflect those that might be found in real data and to facilitate comparisons with previous studies. Test length was not manipulated; 40 item tests were constructed as in Narayanan and Swaminathan (1994), Narayanan and Swaminathan (1996), and Rogers and Swaminathan (1993) to represent a short but reliable standardized achievement test.

First, sample size is an important factor in any examination of power. Type I error and power rates for MH, SIB, and LR DIF procedures increase as the sample size of the reference and focal groups increase when only significance tests are used (Narayanan & Swaminathan 1994; Rogers & Swaminathan, 1993; Roussos & Stout, 1996b). Three reference group sample sizes (N_R=250, N_R=500, N_R=1000) were crossed with three focal group sample sizes (N_F=250, N_F=500, N_F=1000) with the restriction that $N_R \geq N_F$ to produce six sample size combinations.

Second, ability distribution differences were considered. Although several studies have demonstrated adherence to nominal Type I error rates with ability differences as large as one standard deviation between the reference and focal groups (Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993), ability differences are a common phenomenon and can have significant interactions with other variables. Hence, two levels were considered in this study. In the equal ability distribution condition, both the reference and focal group abilities were randomly generated to form normal distributions with mean 0.0 and standard deviation 1.0. In the unequal ability distribution condition, the focal group ability was modified to a normal distribution with mean -0.5 and standard deviation 1.0.

Third, the percentage of items containing DIF was a factor of interest. Because the percentage of DIF items can reduce the validity of the matching variable, it is also expected to affect Type I error and power rates. Three levels were considered: 10% of the items favoring the reference group, 20% of the items favoring the reference group, and 20% of the items containing DIF with 10% favoring the reference and 10% favoring the focal group (e.g., four, eight, and eight items, respectively). The first two conditions in which all the items favor the reference group were designed to represent the situation where some form of systematic difference is present in the DIF items. Such a situation might be expected, for example, in gender comparisons where females consistently outperform males. The third condition, which will be referred to as 10%/10% DIF, is intended to reflect situations where differences would be random rather than systematic. In situations such as test translation, there is often insufficient a priori evidence to suggest a systematic reason for items to favor only one group because translation errors tend to be random.

Thus, DIF analyses for datasets with six levels of sample size, two levels of ability distributions, and three levels of DIF item percentages were fully crossed for 36 conditions. Each condition was replicated 100 times to facilitate Type I error and power calculations.

Data Generation and Analysis

The three parameter logistic item response model was used for the generation of examinee response data which necessitated the stipulation of item parameters for both the non-DIF and DIF items. The non-DIF items included in each test were randomly selected from an administration of the Graduate Management Admissions Test, as cited in Narayanan and Swaminathan (1994). The same item parameters were used for both the reference and focal groups resulting in unbiased items that were expected to reflect realistic items that were free of DIF. These item parameters are shown in Table 1. The first 32 items were used for the non-DIF items needed in the 20% and 10%/10% conditions and all 36 items in the 10% DIF condition.

The DIF items included in each test were designed to reflect those which may be found in standardized tests. Several characteristics of these items should be highlighted and are included with item parameters in Table 2. Items with a range of discriminations and difficulties were included, as were uniform and nonuniform items. For each test, the ratio of uniform to nonuniform DIF items was kept at 3:1 to reflect the more frequent occurrence of uniform DIF. Furthermore, DIF effect sizes based on the area between item response functions (Raju, 1988) were limited to 0.4 and 0.6 to reflect DIF of moderate size. Larger DIF effect sizes were not simulated because they have been shown to be flagged with high frequency by MH, SIB, and LR DIF procedures (Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993). In addition, large DIF is infrequently found in practice (Linn, 1993). The first four items were used in the 10% DIF percentage condition. All eight items were used in the 20% and 10%/10% DIF conditions, with the focal and reference group parameters for items 3, 6, 7, and 8 interchanged for the 10%/10% DIF condition.

For each condition, the probability of a correct response for an item given the three parameter logistic model was calculated by substituting the appropriate item and ability parameters for each examinee. The item was scored correct if the probability of a correct response exceeded a random number from a uniform distribution in the interval [0, 1] and incorrect otherwise. This was completed with an Excel macro.

In computing the chi-square and $R^2\Delta$ values, the sum of the test items was used for the matching variable. Furthermore, in order to avoid collinearity problems between the interaction term and the other explanatory variables, θ and g were centered through z-score transformations before creating the interaction term θg (see Zumbo & Thomas, 1996, p. 3). LR statistics were computed with SPSS.

To enable comparisons between the two degree of freedom chi-square test and the separate one degree of freedom chi-square tests which included the effect size measures $R^2\Delta$-U and $R^2\Delta$-N, the following definitions were applied for Type I error and power. For the two degree of freedom chi-square test, a DIF item was correctly identified if the probability of the two degree of freedom test was less than 0.05 and a Type I error occurred if a non-DIF item was similarly flagged. For the separate one degree of freedom chi-square tests, an DIF item was correctly identified if the probability of either one degree of freedom test was less than 0.05 and the corresponding $R^2\Delta$ was greater than or equal to 0.035. Similarly, a Type I error occurred if a non-DIF item met this condition.

Results

The results of the simulation study are presented below for the 10%, 20%, and 10%/10% conditions. In each condition, results for the equal ability distributions are presented before the unequal ability distributions with Type I error discussions preceding overall, uniform, and nonuniform power discussions, respectively.

10% DIF

Table 3 displays the Type I error and power frequencies and percentages for the 10% DIF condition with equal ability distributions for both the two degree of freedom chi-square test and the separate one degree of freedom chi-square tests with $R^2\Delta$ classification method using the rules described above.

The Type I errors rate for the two degree of freedom chi-square test exceeded the nominal level for all sample sizes and increased from 5.3% to 7.3% for the N_R=250, N_F=250 and N_R=1000, N_F=1000 conditions, respectively. The opposite trend was observed when $R^2\Delta$ was used. Type I errors decreased as sample size increased with a 1.0% Type I error rate in the N_R=1000, N_F=1000 condition. This corresponds to 35 false positives of 3600 non-DIF items.

The overall power of both procedures increased as sample size increased with the exception of the N_R=1000, N_F=250 condition when the $R^2\Delta$ procedure was employed. The power of the $R^2\Delta$ procedure for this unbalanced condition was lower than the N_R=500, N_F=500 condition. Similarly, the power to detect uniform DIF increased as sample sizes increased with the exception of the most unbalanced design using the $R^2\Delta$ procedure. The low power for the N_R=1000, N_F=250 sample size was problematic across DIF percentage and ability distributions often resulting in lower power than the N_R=500,

N_F=250 condition. At present, no explanation is available for this outcome. The power to detect nonuniform DIF increased from 19% to 82% for the two degree of freedom chi-square test whereas sample size made little difference in the power of the $R^2\Delta$ procedure that varied between 23% and 35%. Finally, with the exception of only the N_R=250, N_F=250 condition, the power of the two degree of freedom chi-square test exceeded that of the $R^2\Delta$ procedure. For example, in the N_R=1000, N_F=1000 condition the overall power of the two degree of freedom chi-square test was 95.5% compared to 82.3% for the $R^2\Delta$ procedure. However, this result needs to be tempered by considering the large differences in the Type I error rates when larger sample sizes are used.

Although the inflated Type I rates have been discussed in previous studies, a comparison of the Type I error and power frequencies more clearly demonstrates the problem. For the N_R=1000, N_F=1000 condition, 264 of the 646 items identified as containing DIF were false positives for the two degree of freedom chi-square test as compared to 35 of 364 for the $R^2\Delta$ procedure. That is, nearly 41% of the items flagged as containing DIF were Type I errors when only the two degree of freedom chi-square test was used. Normally, large sample sizes and a low percentage of items containing DIF, as simulated in this condition, is considered ideal. However, the ratio of Type I errors to items flagged as containing DIF is highest with large sample sizes and when the percentage of items containing DIF is low. Clearly, consideration of both Type I error and power frequencies are essential to balance high identification of DIF items and efficient use of resources.

Table 4 shows the Type I error and power results for the 10% DIF with unequal ability distributions condition. As in the previous condition, Type I errors increased as sample sizes increased but the Type I error rate was higher. In contrast, the overall, uniform, and nonuniform power were lower in the unequal ability distribution. However, power trends were consistent with those described in the 10% DIF with equal ability distribution condition.

<u>20% DIF</u>

Tables 5 and 6 show the results for the 20% DIF with equal and unequal ability distribution conditions, respectively. As with the 10% DIF conditions, Type I errors increased as sample size increased for the two degree of freedom chi-square procedure and decreased for the $R^2\Delta$ procedure. Still, the Type I error rate was higher for comparable sample sizes and ability distributions for both conditions peaking at 15.8% for the $N_R=1000$, $N_F=1000$ unequal ability distribution condition with the two degree of freedom chi-square test.

The trend of power increasing with sample size was comparable to the 10% DIF conditions. However, a comparison of power for the 10% and 20% DIF conditions is complicated by the different properties of the items containing DIF in the 20% condition. Uniform DIF power is generally lower for the 20% DIF conditions which may be partially due to the inclusion of relatively easy and difficult items. Item characteristics such as difficulty and discrimination have been demonstrated to effect the power of DIF procedures. Similarly, nonuniform DIF detection generally improved for the 20% DIF conditions. This outcome is likely due to the inclusion of a second nonuniform DIF item with a larger effect size (i.e., area between item response function equal to 0.6) and not related to the percentage of DIF items present. A comparison of the power of those items included in both the 10% and 20% DIF conditions indicates superior power in the 10%

DIF condition. This supports the findings reported by Rogers and Swaminathan (1993) and Narayanan and Swaminathan (1996).

10/10% DIF

Tables 7 and 8 show the results for the 10%/10% DIF with equal and unequal ability distribution conditions, respectively. Type I error and power trends were consistent with those reported in the 10% and 20% DIF conditions. Generally, the power was higher and the Type I error rates were lower for the 10%/10% DIF condition than either the 10% or 20% DIF conditions for comparable sample sizes and ability distributions. Two hypotheses for this interesting result seem reasonable. First, the bias present in the individual DIF items when aggregated would result, overall, in an approximately unbiased test. In turn, this would provide a superior conditioning variable. Second, the balancing of items favoring both the focal and reference groups may benefit from the compensatory nature of current DIF procedures which typically have effect sizes that sum to approximately zero across items (Camilli, 1993; Williams, 1997).

Discussion and Conclusion

The first purpose of this study was the systematic comparison of $R^2\Delta$ to existing measures and the development of negligible, moderate, and large DIF classification guidelines based on this comparison. Visual representation of data from a large-scale testing program suggests a consistent curvilinear relationship between $R^2\Delta$-U and the SIB effect size measure $\hat{\beta}_U$. The cubic regression procedure resulted in guidelines very different from those suggested by Zumbo and Thomas (1996). This contrast is intriguing in light of the ongoing difficulties in linking statistical and substantive DIF reviews (e.g., Camilli & Shepard, 1994; Linn, 1993; O'Neill & McPeek, 1993, Roussos & Stout, 1996a; Willingham & Cole, 1997).

Cohen (1988) defined small, medium, and large effect sizes as follows: A small effect size is noticeably smaller than medium but not so small as to be necessarily trivial, a medium effect size is likely to be visible to the naked eye of a careful observer, and a large effect size is the same distance above medium as small is below it. The effect sizes used in this study, as modeled by item response functions, meet, at minimum, the criteria for a medium effect size. Recall, for uniform DIF items, difficulty parameters differed by 0.50 and 0.75 while for nonuniform DIF items discrimination parameters differed by 0.35 and 0.85 for the reference and focal groups. Yet by the classification methods outlined by Zumbo and Thomas (1996), only 6.8% of the DIF items in this study would be identified as containing moderate DIF. In contrast, these same items were identified as containing at least moderate DIF by the criteria derived from those in use with SIB 68.2% of the time. On the other hand if an interpretation similar to other R^2 measures is applied, only 3.5% of the variation explained by the grouping variable would trigger the moderate

DIF classification. Some social scientists would consider this effect small rather than moderate or large. Thus, two alternative statistical approaches suggest notably different perspectives on the expectation to substantively explain the underlying psychology of statistically flagged DIF. That is, differences in item response functions we hope to interpret substantively are flagged by effect sizes that, in turn, account for a small proportion of variance in the dependent variable. The juxtaposition of these two statistical interpretations may provide a partial explanation for the current discontinuity between statistical and substantive interpretations and merits continued investigation.

The second purpose of this study was to consider the effects the inclusion of $R^2\Delta$ had on both Type I error and power rates of the LR DIF procedure. To investigate this approach a simulation study considering sample size, ability distribution differences, and percentage of items containing DIF was conducted. Type I errors decreased as sample size increased when $R^2\Delta$ was used and generally were below the nominal alpha level of .05 when the combined sample size exceeded 1000. When small sample sizes were used, Type I errors were above the nominal level.

In general, power decreased with the inclusion of $R^2\Delta$ although with larger sample sizes the benefit of reduced Type I errors may supercede the loss of power. The power to detect uniform DIF seems reasonable at 75.3% across all 36 conditions although sample size needs to be considered. This finding is similar to Rogers and Swaminathan (1993) who reported 73% power for uniform DIF on 40 item tests. The power to detect nonuniform DIF was much lower at 32.5% across all conditions. Initially, an inappropriate classification value for $R^2\Delta$-N was believed to produce this result. However, further analysis of the source of Type I errors indicated approximately half the errors were a result of uniform DIF being flagged and the other half the result of nonuniform DIF being flagged (i.e., 2740 uniform; 2798 nonuniform; 89 both uniform and nonuniform). This suggests the classification scores for $R^2\Delta$-N are appropriate. However, only two nonuniform DIF items were considered in this study, therefore, this classification guideline warrants additional study and comparison to the effect size measure associated with the Crossing Simultaneous Item Bias Test. Furthermore, this outcome suggests if only uniform DIF were screened Type I errors would be approximately half of those reported in this study.

Finally, for both uniform and nonuniform DIF, the N_R=1000, N_F=250 condition was problematic resulting in unusually low power. Approximately equal sample sizes should be used to alleviate this concern whenever possible. In addition, future research should further investigate the effects of unequal reference and focal group sample sizes on DIF detection.

In conclusion, it warrants repeating that an inclusive view of the variables associated with statistical inferences is required in DIF. Sample size, Type I error rate, power, and effect sizes are intertwined and need to be considered together with careful attention to the inferences, and their consequences, drawn from a statistical test. Future research comparing statistical approaches to identify DIF must include effect size measures with attention to both Type I error and power. These considerations will improve the generalizability of results from simulation studies to practice.

Endnotes

[1] Because the $R^2\Delta$ effect size measure may be additively partitioned, it need not consider explanatory variables in isolation. Indeed, Zumbo and Thomas (1996) suggested considering the uniform and nonuniform terms simultaneously. The separate uniform and nonuniform approach is based on the assumption that this specific information may assist content reviewers in interpreting DIF.

[2] Each data set was also fit separately with a cubic regression model to develop a regression equation. The SIB cutscores were substituted into these regression equations to determine comparable cut scores. The mean of these values, across the four data sets, corresponded to cutscores of 0.036 and 0.070. The values for the combined data set were 0.035 and 0.069. Given the similarity of the two approaches only one is reported in text.

Table 1

Item Parameters for Non-DIF Items

Item	A	B	C	Item	A	B	C
1	0.44	-0.30	0.20	19	0.55	1.09	0.20
2	0.55	-1.06	0.20	20	1.40	1.64	0.20
3	0.82	1.02	0.20	21	0.92	1.13	0.20
4	0.52	-1.96	0.20	22	0.64	-1.55	0.20
5	1.02	1.28	0.20	23	1.01	0.81	0.20
6	0.82	0.61	0.20	24	0.61	-0.53	0.20
7	0.92	0.42	0.20	25	0.70	1.05	0.20
8	0.65	1.68	0.20	26	1.02	0.64	0.20
9	0.56	-2.70	0.20	27	0.48	2.12	0.20
10	0.29	-1.39	0.20	28	1.01	0.91	0.20
11	0.35	-1.12	0.20	29	0.53	0.87	0.20
12	0.31	-1.37	0.20	30	0.36	-2.63	0.20
13	1.05	0.10	0.20	31	1.12	-1.21	0.20
14	0.51	-0.09	0.20	32	0.86	-0.57	0.20
15	0.73	0.61	0.20	33	0.59	-1.29	0.20
16	0.88	0.95	0.20	34	0.56	0.40	0.20
17	1.11	-0.35	0.20	35	1.09	1.11	0.20
18	1.32	0.57	0.20	36	0.88	-0.93	0.20

Note. A, B, and C correspond to the discrimination, difficulty, and pseudo-guessing parameters in the 3-PL IRT model.

Table 2

Item Parameters Used to Generate DIF Items

Item	A_R	B_R	C_R	A_F	B_F	C_F	Area	DIF Type
1	1.25	-0.25	0.20	1.25	0.25	0.20	0.40	Uniform
2	0.50	-0.38	0.20	0.50	0.38	0.20	0.60	Uniform
3	0.90	-1.63	0.20	0.90	-0.88	0.20	0.60	Uniform
4	0.45	0.00	0.20	0.79	0.00	0.20	0.60	Nonuniform
5	0.90	1.00	0.20	0.90	1.50	0.20	0.40	Uniform
6	1.25	0.88	0.20	1.25	1.63	0.20	0.60	Uniform
7	0.90	-0.25	0.20	0.90	0.25	0.20	0.40	Uniform
8	0.80	0.00	0.20	1.65	0.00	0.20	0.40	Nonuniform

Note. A, B, and C correspond to the discrimination, difficulty, and pseudo-guessing parameters in the 3-PL IRT model.

Table 3

Frequencies and Percentages for LR and LR with $R^2\Delta$ for 10% DIF Equal Ability Distribution Condition

		Frequency		Percentage	
		χ^2_2	$R^2\Delta$	χ^2_2	$R^2\Delta$
N_R=250, N_F=250	Overall power	237	249	59.3	62.3
	Uniform power	218	226	72.7	75.3
	Nonuniform power	19	23	19.0	23.0
	Type I error	190	269	5.3	7.5
N_R=500, N_F=250	Overall power	270	265	67.5	66.3
	Uniform power	244	237	81.3	79.0
	Nonuniform power	26	28	26.0	28.0
	Type I error	219	199	6.1	5.5
N_R=500, N_F=500	Overall power	331	300	82.8	75.0
	Uniform power	288	265	96.0	88.3
	Nonuniform power	43	35	43.0	35.0
	Type I error	189	119	5.3	3.3
N_R=1000, N_F=250	Overall power	317	253	79.3	63.3
	Uniform power	269	227	89.7	75.7
	Nonuniform power	48	26	48.0	26.0
	Type I error	204	94	5.7	2.6
N_R=1000, N_F=500	Overall power	356	303	89.0	75.8
	Uniform power	296	273	98.7	91.0
	Nonuniform power	60	30	60.0	30.0
	Type I error	255	73	7.1	2.0
N_R=1000, N_F=1000	Overall power	382	329	95.5	82.3
	Uniform power	300	295	100.0	98.3
	Nonuniform power	82	34	82.0	34.0
	Type I error	264	35	7.3	1.0
Across sample size	Overall power	316	283	79.0	70.8
	Uniform power	269	254	89.7	84.6
	Nonuniform power	46	29	46.0	29.0
	Type I error	220	132	6.1	3.7

Table 4

Frequencies and Percentages for LR and LR with $R^2\Delta$ for 10% DIF Unequal Ability Distribution Condition

		Frequency		Percentage	
		χ^2_2	$R^2\Delta$	χ^2_2	$R^2\Delta$
N_R=250, N_F=250	Overall power	253	250	63.3	62.5
	Uniform power	232	227	77.3	75.7
	Nonuniform power	21	23	21.0	23.0
	Type I error	252	379	7.0	10.5
N_R=500, N_F=250	Overall power	274	258	68.5	64.5
	Uniform power	249	235	83.0	78.3
	Nonuniform power	25	23	25.0	23.0
	Type I error	304	266	8.4	7.4
N_R=500, N_F=500	Overall power	313	292	78.3	73.0
	Uniform power	268	261	89.3	87.0
	Nonuniform power	45	31	45.0	31.0
	Type I error	313	191	8.7	5.3
N_R=1000, N_F=250	Overall power	304	231	76.0	57.8
	Uniform power	270	215	90.0	71.7
	Nonuniform power	34	16	34.0	16.0
	Type I error	327	160	9.1	4.4
N_R=1000, N_F=500	Overall power	337	280	84.3	70.0
	Uniform power	293	261	97.7	87.0
	Nonuniform power	44	19	44.0	19.0
	Type I error	383	113	10.6	3.1
N_R=1000, N_F=1000	Overall power	367	309	91.8	77.3
	Uniform power	296	278	98.7	92.7
	Nonuniform power	71	31	71.0	31.0
	Type I error	471	71	13.1	2.0
Across sample size	Overall power	308	253	77.0	63.3
	Uniform power	268	229	89.3	76.4
	Nonuniform power	40	24	40.0	24.0
	Type I error	342	197	9.5	5.5

Table 5

Frequencies and Percentages for LR and LR with $R^2\Delta$ for 20% DIF Equal Ability Distribution Condition

		Frequency		Percentage	
		χ^2_2	$R^2\Delta$	χ^2_2	$R^2\Delta$
N_R=250, N_F=250	Overall power	437	454	54.6	56.8
	Uniform power	362	374	60.3	62.3
	Nonuniform power	75	80	37.5	40.0
	Type I error	231	317	7.2	9.9
N_R=500, N_F=250	Overall power	527	450	65.9	56.3
	Uniform power	427	372	71.2	62.0
	Nonuniform power	100	78	50.0	39.0
	Type I error	238	232	7.4	7.3
N_R=500, N_F=500	Overall power	653	525	81.6	65.6
	Uniform power	522	439	87.0	73.2
	Nonuniform power	131	86	65.5	43.0
	Type I error	262	166	8.2	5.2
N_R=1000, N_F=250	Overall power	592	384	74.0	48.0
	Uniform power	478	343	79.7	57.2
	Nonuniform power	114	41	57.0	20.5
	Type I error	264	98	8.3	3.1
N_R=1000, N_F=500	Overall power	715	516	89.4	64.5
	Uniform power	552	432	92.0	72.0
	Nonuniform power	163	84	81.5	42.0
	Type I error	289	76	9.0	2.4
N_R=1000, N_F=1000	Overall power	771	582	96.4	72.8
	Uniform power	588	491	98.0	81.8
	Nonuniform power	183	91	91.5	45.5
	Type I error	336	47	10.5	1.5
Across sample size	Overall power	616	485	77.0	60.6
	Uniform power	488	409	81.4	68.1
	Nonuniform power	128	77	63.8	38.3
	Type I error	270	156	8.4	4.9

Table 6

Frequencies and Percentages for LR and LR with $R^2\Delta$ for 20% DIF Unequal Ability Distribution Condition

		Frequency		Percentage	
		χ^2_2	$R^2\Delta$	χ^2_2	$R^2\Delta$
N_R=250, N_F=250	Overall power	415	421	51.9	52.6
	Uniform power	369	372	61.5	62.0
	Nonuniform power	46	49	23.0	24.5
	Type I error	229	314	7.2	9.8
N_R=500, N_F=250	Overall power	482	397	60.3	49.6
	Uniform power	418	351	69.7	58.5
	Nonuniform power	64	46	32.0	23.0
	Type I error	272	272	8.5	8.5
N_R=500, N_F=500	Overall power	597	481	74.6	60.1
	Uniform power	511	429	85.2	71.5
	Nonuniform power	86	52	43.0	26.0
	Type I error	364	219	11.4	6.8
N_R=1000, N_F=250	Overall power	513	315	64.1	39.4
	Uniform power	445	289	74.2	48.2
	Nonuniform power	68	26	34.0	13.0
	Type I error	333	140	10.4	4.4
N_R=1000, N_F=500	Overall power	670	451	83.8	56.4
	Uniform power	547	405	91.2	67.5
	Nonuniform power	123	46	61.5	23.0
	Type I error	436	143	13.6	4.5
N_R=1000, N_F=1000	Overall power	739	508	92.4	63.5
	Uniform power	577	463	96.2	77.2
	Nonuniform power	162	45	81.0	22.5
	Type I error	506	81	15.8	2.5
Across sample size	Overall power	569	429	71.2	53.6
	Uniform power	478	385	79.6	64.1
	Nonuniform power	92	44	45.8	22.0
	Type I error	357	195	11.1	6.1

Table 7

Frequencies and Percentages for LR and LR with $R^2\Delta$ for 10%/10% DIF Equal Ability Distribution Condition

		Frequency		Percentage	
		χ^2_2	$R^2\Delta$	χ^2_2	$R^2\Delta$
N_R=250, N_F=250	Overall power	508	543	63.5	67.9
	Uniform power	438	457	73.0	76.2
	Nonuniform power	70	86	35.0	43.0
	Type I error	182	261	5.7	8.2
N_R=500, N_F=250	Overall power	599	544	74.9	68.0
	Uniform power	497	465	82.8	77.5
	Nonuniform power	102	79	51.0	39.5
	Type I error	164	162	5.1	5.1
N_R=500, N_F=500	Overall power	685	603	85.6	75.4
	Uniform power	559	517	93.2	86.2
	Nonuniform power	126	86	63.0	43.0
	Type I error	137	104	4.3	3.3
N_R=1000, N_F=250	Overall power	661	497	82.6	62.1
	Uniform power	541	448	90.2	74.7
	Nonuniform power	120	49	60.0	24.5
	Type I error	170	76	5.3	2.4
N_R=1000, N_F=500	Overall power	735	617	91.9	77.1
	Uniform power	585	531	97.5	88.5
	Nonuniform power	150	86	75.0	43.0
	Type I error	159	48	5.0	1.5
N_R=1000, N_F=1000	Overall power	773	666	96.6	83.3
	Uniform power	596	563	99.3	93.8
	Nonuniform power	177	103	88.5	51.5
	Type I error	163	30	5.1	0.9
Across sample size	Overall power	660	578	82.5	72.3
	Uniform power	536	497	89.3	82.8
	Nonuniform power	124	82	62.1	40.8
	Type I error	163	114	5.1	3.5

Table 8

Frequencies and Percentages for LR and LR with $R^2\Delta$ for 10%/10% DIF Unequal Ability Distribution Condition

		Frequency		Percentage	
		χ^2_2	$R^2\Delta$	χ^2_2	$R^2\Delta$
N_R=250, N_F=250	Overall power	491	504	61.4	63.0
	Uniform power	423	432	70.5	72.0
	Nonuniform power	68	72	34.0	36.0
	Type I error	214	306	6.7	9.6
N_R=500, N_F=250	Overall power	561	502	70.1	62.8
	Uniform power	460	424	76.7	70.7
	Nonuniform power	101	78	50.5	39.0
	Type I error	205	195	6.4	6.1
N_R=500, N_F=500	Overall power	684	580	85.5	72.5
	Uniform power	546	489	91.0	81.5
	Nonuniform power	138	91	69.0	45.5
	Type I error	227	149	7.1	4.7
N_R=1000, N_F=250	Overall power	614	434	76.8	54.3
	Uniform power	499	387	83.2	64.5
	Nonuniform power	115	47	57.5	23.5
	Type I error	228	102	7.1	3.2
N_R=1000, N_F=500	Overall power	734	565	91.8	70.6
	Uniform power	576	469	96.0	78.2
	Nonuniform power	158	96	79.0	48.0
	Type I error	253	72	7.9	2.3
N_R=1000, N_F=1000	Overall power	780	637	97.5	79.6
	Uniform power	595	533	99.2	88.8
	Nonuniform power	185	104	92.5	52.0
	Type I error	285	48	8.9	1.5
Across sample size	Overall power	644	537	80.5	67.1
	Uniform power	517	456	86.1	75.9
	Nonuniform power	128	81	63.8	40.7
	Type I error	235	145	7.4	4.5

Figure 1. Typical uniform DIF item

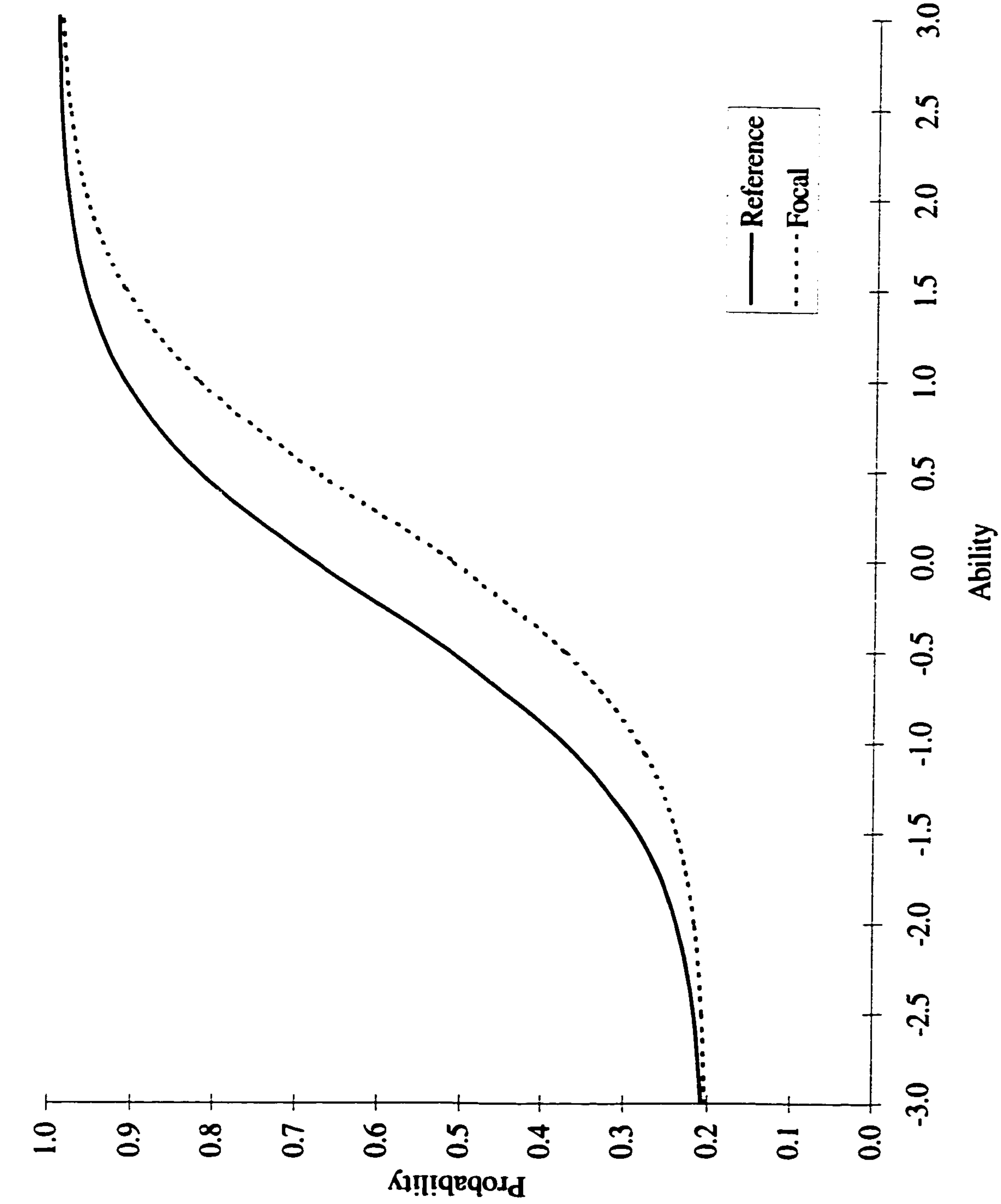


Figure 2. Typical nonuniform DIF item

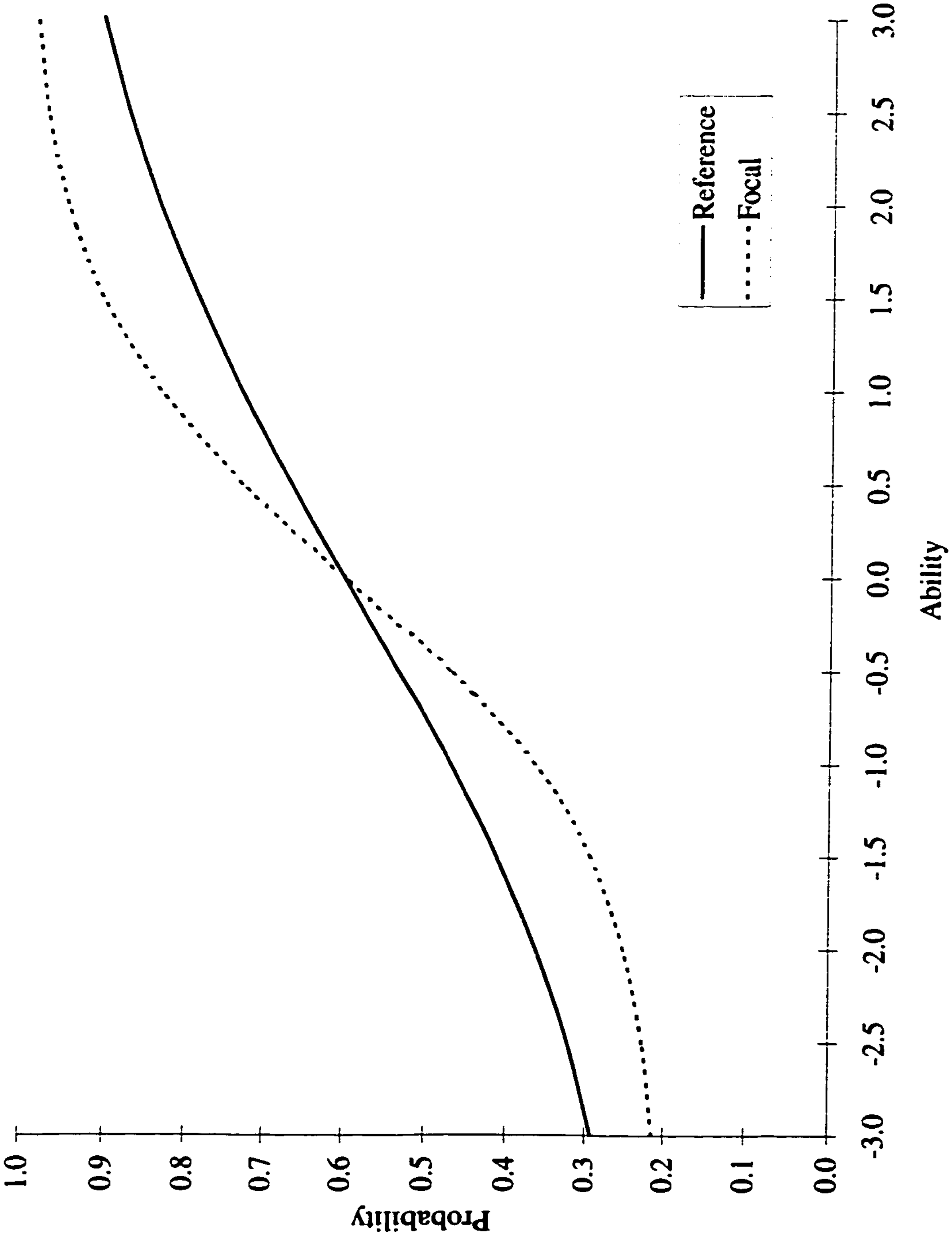
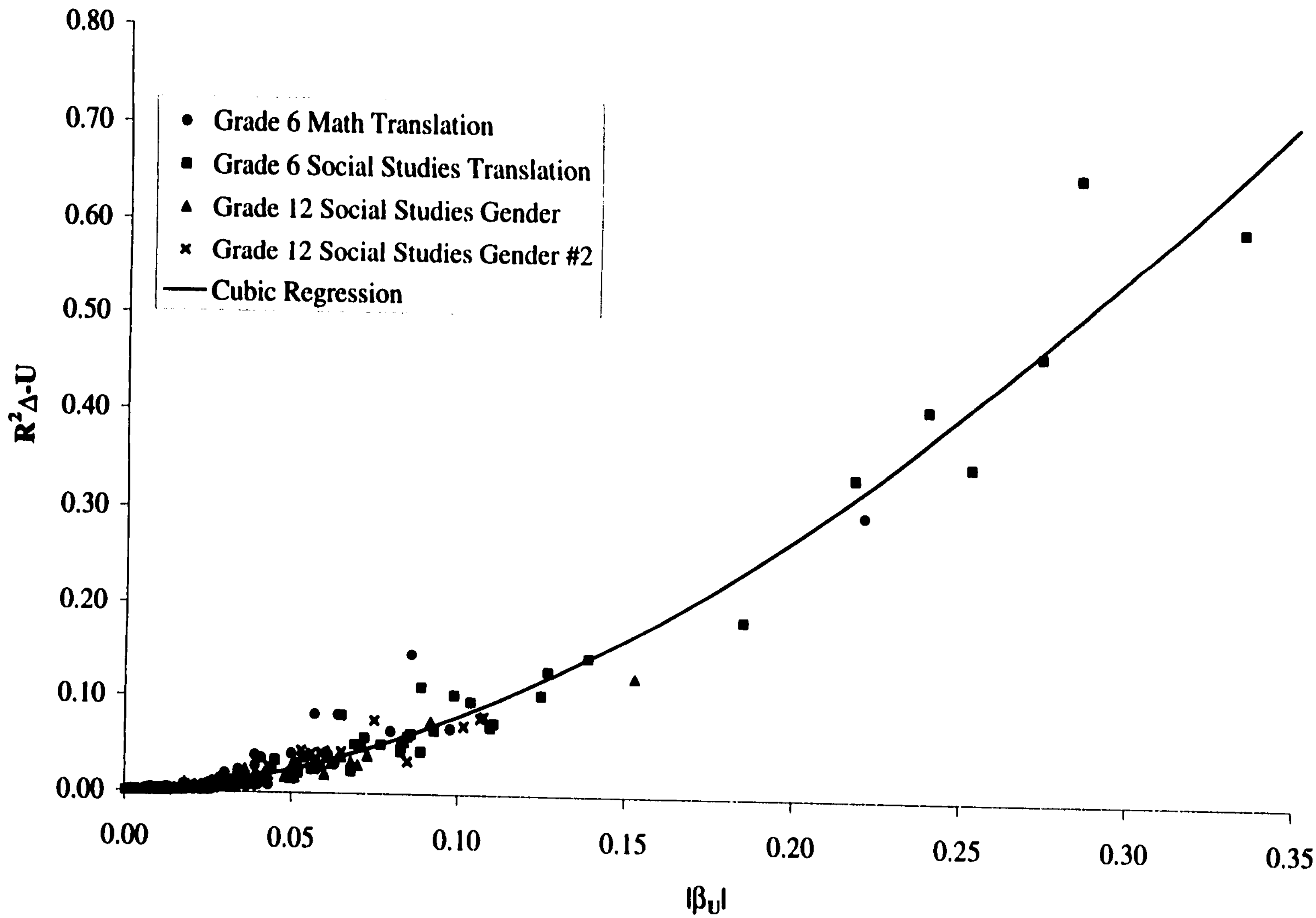


Figure 3. Scatterplots for four large-scale achievement tests and cubic regression curve predicting $R^2\Delta$ from β_U



Bibliography

Camilli, G., (1993). The case against item bias techniques based on internal criteria: Do item bias procedures obscure test fairness issues? The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland, & H. Wainer (Eds.), Differential item functioning (pp. 397-413). Hillsdale, NJ: Erlbaum.

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased items. Thousand Oaks, CA: Sage Publications.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. Educational Measurement: Issues and Practice, 17, 31-44.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Cohen, J. (1994). The earth is round (p<.05). American Psychologist, 49, 997-1993.

Gierl, M. J., Rogers, W. T., & Klinger, D. A. (1999). Using statistical judgement reviews to identify and iterpret translation DIF. Paper presented at the Annual Meeting of the National Council for Measurement in Education, Montreal, QC, Canada.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Li, H., & Stout, W. (1996). A new procedure for detection of crossing DIF. Psychometrika, 61, 647-677.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland, & H. Wainer (Eds.), Differential item functioning (pp. 349-366). Hillsdale, NJ: Erlbaum.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement, 17, 297-334.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and Simultaneous Item Bias Procedures for detecting differential item functioning. Applied Psychological Measurement, 18, 315-328.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. Applied Psychological Measurement, 20, 257-274.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland, & H. Wainer (Eds.), Differential item functioning (pp. 255-276). Hillsdale, NJ: Erlbaum.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. Applied Psychological Measurement, 19, 23-37.

Pratt, J. W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained, in T. Pukkila and S. Puntanen (eds.), Proceedings of the Second International Conference in Statistics. Tampere, Finland. p. 245-260.

Pregibon, D. (1981). Logistic regression diagnostics. Annals of Statistics, 9, 705-724.

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

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

Roussos, L. A., & Stout, W. F. (1996a). A multidimensionality-based DIF analysis paradigm. Applied Psychological Measurement, 20, 355-371.

Roussos, L. A., & Stout, W. F. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. Journal of Educational Measurement, 33, 215-230.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/dif from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometrika, 58, 159-194.

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

Williams, V. S. L., (1997). The "unbiased" anchor: Bridging the gap between DIF and item bias. Applied Measurement in Education, 3, 253-267.

Willingham, W. W., & Cole, N. S. (1997). Gender and fair assessment. Hillsdale, NJ: Erlbaum.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modelling as a unitary framework for binary and likert-type (ordinal) item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D., & Thomas, D. R. (1996, October). A measure of DIF effect size using logistic regression procedures. Paper presented at the National Board of Medical Examiners, Philadelphia, PA.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. Journal of Educational Measurement, 26, 55-66.