ACT Research Report Series 94-1

Identifying Nonuniform DIF in Polytomously Scored Test Items

Judith Spray
Tim Miller

June 1994

ACT

Identifying Nonuniform DIF in Polytomously Scored Test Items

Judy Spray
Tim Miller

Abstract

Computer simulations under three conditions of polytomous DIF compared the ability of three different statistical procedures to detect nonuniform DIF. The procedures were a nominal and an ordinal extension of the Mantel-Haenszel statistic, and logistic discriminant function analysis. Results showed that only the logistic discriminant function analysis could detect all types of nonuniform DIF simulated when sample sizes were moderate-to-large (i.e. $N > 500$). This procedure is recommended when nonuniform DIF identification is required.

Identifying Nonuniform DIF in Polytomously Scored Test Items

The use of polytomously scored items in addition to, or in place of the more traditional correct/incorrect item formats, requires reconsideration of some of the psychometric procedures that are specific to the dichotomous situation. In particular, the identification of differential item functioning or DIF within each of J categories of a polytomously scored item requires either modifications of procedures that are currently used for dichotomous items, or the creation of new procedures that are especially suited for multiple-category item scoring. Several extensions of the existing Mantel-Haenszel procedure, a popular method for identifying DIF in dichotomous items, have been suggested for the polytomous case. These extended Mantel-Haenszel procedures are similar to those used in the dichotomous situation for 0/1 item responses which have been tabulated in a 2 X 2 X K table, in that they assume that there is no three-way interaction. In other words, nonuniform DIF is assumed not to exist. The only way that this assumption can be tested is if a procedure is used that allows for a specific test of the presence of the three-way interaction. Examples include tests of significance of the interaction term in the fitting of a log-linear model, or of the interaction coefficient in a logistic regression model (Swaminathan & Rogers, 1990).

The identification of nonuniform DIF might be more important in a polytomous item than in a dichotomous one because there are potentially more ways in which the group-by-response-by-score interaction can manifest itself in the polytomous situation. For example, it is possible that in addition to the usual nonuniform DIF situation in which the proportion of examinees in a group with some response, $U = u$, varies as a function of the conditioning score, one could have the situation where the proportion remains constant throughout the score scale but reverses group direction for different item response categories. Although this is not the typical way in

which nonuniform DIF occurs, its detection is still important. Any useful polytomous DIF procedure should be powerful enough to detect such occurrences with sufficiently large power.

Another method proposed to detect situations of nonuniform DIF in polytomous items is called *logistic discriminant function analysis* (LDFA). This method has recently been suggested as a useful procedure for the identification of DIF (both uniform and nonuniform cases) in polytomous items (Miller & Spray, 1993). The method is similar to those mentioned previously (i.e., log-linear modeling and logistic regression) in that a separate test of the significance of the interaction is available. However, the LDFA method is much easier to implement than the logistic regression for the polytomous case (Miller & Spray, 1993). The method is identical to some log-linear modeling approaches (Hanson, 1992), but may be easier to interpret because of graphical procedures which can be used *post hoc* to investigate the direction and magnitude of the DIF visually (Miller & Spray, 1993).

Although they lack separate tests of any possible interaction, several extensions of the Mantel-Haenszel procedure are available for DIF identification in polytomous items, depending upon whether the responses can be treated as nominal or ordinal. Mantel and Haenszel (1959) extended the 2 X 2 X K situation to the 2 X 3 X K case with 3 nominal levels of response, and showed that a summary chi-squared statistic with 2 degrees of freedom could be obtained (pp. 743-745). The authors also gave approximations for the more general, 2 X J X K situation, where there are J nominal response levels. Agresti (1990) later summarized work which gave exact, rather than approximate, procedures for the more general I X J X K case.

Mantel (1963) later proposed an extension whereby the J responses are scored or weighted by ordered scores. Mantel showed that the summary score statistic was simply the weighted sum of the J frequencies, weighted by the J scores at each of the K levels. This amounted to testing

a null hypothesis about the *mean* level of the J responses, so that the summary statistic was tested with a single degree of freedom only (Mantel, 1963). This score statistic was extended by Landis, Heyman, and Koch (1978) to the I X J X K situation where either the J response levels, the I levels, or both are ordered and ordinal scores can be assigned to the responses. A convenient vector representation of this situation is provided by Agresti (1990, p. 286).

In the 2 X 2 X K situation with dichotomous items, the Mantel-Haenszel procedure often is quite robust in detecting DIF, even when there is a serious violation of the assumption of no three-way interaction. Therefore, the purpose of this paper was to report a series of computer simulations in which different types of DIF were present in simulated polytomous item responses. Three procedures were then used to detect the presence of DIF. The procedures were compared on the basis of their ability to detect true DIF when it existed (i.e., statistical power) and to detect it when it did *not* exist (i.e., Type I error). The procedures used in the simulations were (1) the extended Mantel-Haenszel test on nominal data with J-1 degrees of freedom, (2) the Mantel score statistic on ordinal data with one degree of freedom , and (3) the LDFA procedure. Each procedure is briefly described below.

Logistic Discriminant Function Analysis

The logistic discriminant function, which is estimated via the LDFA procedure, can be written as

$$\text{Prob}(G \mid X,U) = \frac{e^{(1-G)(-\alpha_0-\alpha_1 X-\alpha_2 U-\alpha_3 X*U)}}{1 + e^{(-\alpha_0-\alpha_1 X-\alpha_2 U-\alpha_3 X*U)}}, \qquad \textbf{(1)}$$

where the α_i, i = 0, 1, 2, 3, are the discriminant function coefficients to be estimated and G is a Group indicator variable where, for example, G = 1 for the Reference (R) group and G = 0 for the Focal (F) group. U is the item response variable that can take on any one of the J values associated with each item.

Tests of significance of the coefficients, α_3 and α_2, provide answers to the questions concerning nonuniform and uniform DIF, respectively. Specifically, the significance of α_3 is tested by first fitting the hierarchical model given by

$$\text{Prob}(G \mid X,U) = \frac{e^{(1-G)(-\alpha_0-\alpha_1 X-\alpha_2 U)}}{1 + e^{(-\alpha_0-\alpha_1 X-\alpha_2 U)}}. \tag{2}$$

The difference in the log of the likelihood functions obtained from Equations 2 and 1 is used to test for nonuniform DIF or the significance of α_3. The significance of α_2 is tested by next fitting the *null* model, given by

$$\text{Prob}(G \mid X,U) = \text{Prob}(G \mid X) = \frac{e^{(1-G)(-\alpha_0-\alpha_1 X)}}{1 + e^{(-\alpha_0-\alpha_1 X)}}. \tag{3}$$

Equation 3 is termed the *null* model because it represents the probability of group membership only as a function of group sample sizes and group distributions on X. The item response variable is ignored. Thus, the null model given by Equation 3 remains constant from item to item. The difference in the log of the likelihood functions obtained from Equations 3 and 2 is used to test for uniform DIF or the significance of α_2.

Each difference in the log likelihood functions is asymptotically distributed as a chi-squared random variable with one degree of freedom. Thus, with the LDFA procedure, two separate tests can be performed for nonuniform and uniform DIF. The nonuniform DIF test can also be thought of as a test of the no-three-way interaction assumption.

Mantel-Haenszel Extensions

For both extensions described below, the data are assumed to be tabulated in a 2 X J X K table (i.e., 2 groups by J responses by K levels of the conditioning or matching variable).

Nominal Case

The observed counts or absolute frequencies in J-1 cells for the R group across K levels are denoted by $\mathbf{n}_k = (n_{R1k}, n_{R2k}, \ldots, n_{R,J-1,k})'$. The expected frequencies under the hypothesis of conditional independence (i.e., no uniform DIF) are $\mathbf{m}_k = (n_{R+k}n_{+1k}, n_{R+k}n_{+2k}, \ldots, n_{R+k}n_{+,J-1,k})'$. $\mathbf{V}_k$ denotes the null covariance matrix of $\mathbf{n}_k$ (see Agresti, 1990, p.234). Summing over the k strata gives $\mathbf{n} = \Sigma\mathbf{n}_k$, $\mathbf{m} = \Sigma\mathbf{m}_k$, and $\mathbf{V} = \Sigma\mathbf{V}_k$. Then the nominal version of the extended Mantel-Haenszel statistic is given by

$$MH_{\text{nom}} = (\mathbf{n}-\mathbf{m})'\mathbf{V}^{-1}(\mathbf{n}-\mathbf{m}).$$

This statistic has a large-sample chi-squared distribution with J-1 degrees of freedom under the null hypothesis of conditional independence. A significant test implies that uniform DIF is present in the item.

Ordinal Case

The observed counts or absolute frequencies in the kth level are denoted by $\mathbf{n}_k = (n_{R1k}, n_{R2k}, \ldots, n_{R,J,k}, n_{F1k}, n_{F2k}, \ldots, n_{F,J,k})'$. The expected frequencies under the hypothesis of conditional independence (i.e., no uniform DIF) are $\mathbf{m}_k$. $\mathbf{V}_k$ denotes the null covariance matrix of $\mathbf{n}_k$. Also, let $\boldsymbol{U}_k = (u_1,u_2,\ldots,u_J)$, a vector of response category scores, such as $1, 2,\ldots,J$. The scores will usually correspond to the values assigned to the scoring of the item. Then, let $\mathbf{B}_k$ denote a vector of length IJ of score constants, where $\mathbf{B}_k = (u_1,u_2,\ldots,u_J,-u_1,-u_2,\ldots,-u_J)$. The ordinal or scored version of the extended Mantel-Haenszel statistic is given (Agresti, 1990, p. 286) by

$$MH_{\text{ord}} = \{\Sigma\mathbf{B}_k(\mathbf{n}_k-\mathbf{m}_k)\}'\{\Sigma\mathbf{B}_k\mathbf{V}_k\mathbf{B}_k'\}^{-1}\{\Sigma\mathbf{B}_k(\mathbf{n}_k-\mathbf{m}_k)\},$$

where the summation is over k. This statistic has a large-sample chi-squared distribution with 1 degree of freedom under the null hypothesis of conditional independence (i.e., no uniform DIF). A significant test implies that uniform DIF is present in the item. A simpler but equivalent,

algebraic representation of Mantel's score statistic for ordinal responses is given by Mantel (1963, p. 694).

Method

The Simulations

Item responses were generated from Muraki's generalized partial credit model (Muraki, 1991), which gives the item-response density functions or item category characteristic curves (ICCCs) as functions of a unidimensional latent ability, θ. This model can be written as

$$\text{Prob}(U=u_k \mid \theta) = \frac{\exp[\sum_{j=1}^{k} a(\theta - b_j)]}{\sum_{m=1}^{j} \exp[\sum_{j=1}^{m} a(\theta - b_j)]}, \qquad \textbf{(4)}$$

where the b_1, ..., b_j parameters define the points of intersection of the adjoining ICCCs and a represents a slope parameter relating to the discriminating power of the *item.* According to Muraki," ... the discriminating power of each ICCC depends on the combination of the slope and threshold parameters" (p.7). Thus, it is possible to have several different levels of discriminating power for the different item responses within the same test item.

There were 20 items on the simulated tests. Only the last item, item #20, had simulated DIF. The remaining 19 items had identical item parameters for the two groups. These parameters were a = 1.0, b_1 = .00, b_2 = -1.00, b_3 = .5, and b_4 = 1.00. Two sample sizes were used for each group: 500 and 2000. Ability populations were assumed to be identical for both the **focal** (F) and **reference** (R) populations. Ability (i.e., θ) sampling was simulated from a standard normal distribution.

There were three DIF conditions simulated. The first condition was a simple *uniform* DIF case where the a-parameters for both groups remained the same but the b-parameters were shifted or offset by a constant amount. For *Condition 1*, the R group parameters were $\{a=1.0, b_1=0.0, b_2=-1.0, b_3=.5, b_4=1.0\}$, while the F group parameters were $\{a=1.0, b_1=0.0, b_2=-.75, b_3=.75, b_4=1.25\}$. In other words, the item was consistently more difficult for each response category for members of the F group than for comparable members of the R group. Figure 1 illustrates the ICCCs for this item. Response probabilities for the F group are plotted as dotted lines.

see Figure 1 at end of report

For the second condition, nonuniform DIF was simulated where the a-parameters for each group varied but the b-parameters remained the same. For *Condition 2*, the R group parameters were $\{a=1.0, b_1=0.0, b_2=-1.0, b_3=.5, b_4=1.0\}$, while the F group parameters were $\{a=.5, b_1=0.0, b_2=-1.0, b_3=.5, b_4=1.0\}$. This item was more discriminating for the R group for all response categories. See Figure 2.

see Figure 2 at end of report

For the last condition, a less traditional type of nonuniform DIF was simulated. In this case, the a-parameters for each group once again remained the same and only two of the b-parameters varied, but in different directions. For *Condition 3*, the R group parameters were $\{a=1.0, b_1=0.0, b_2=-.75, b_3=.75, b_4=2.0\}$, while the F group parameters were

$\{a=1.0, b_1=0.0, b_2=.75, b_3=-.75, b_4=2.0\}$. This item was therefore easier for the R group for the second category but more difficult for the third category. See Figure 3.

see Figure 3 at end of report

One hundred replications were performed for each of the two sample sizes and for each of the three DIF conditions. A test was significant if the null hypothesis was rejected at a probability level that was less than .05/20 or .0025. Power was computed as the number of replications, out of a possible 100, that a significant test was observed for item # 20. A type I error rate was computed as the number of replications, out of a possible 100, that a significant test was observed for items #1-#19. The summary error rate was the average error rate over those 19 no-DIF items.

Results

The results of the simulations are summarized in Tables 1 and 2. Table 1 gives estimates of power for Item #20 for each of the three different DIF procedures, along with the average chi-squared statistic. For Condition 1, where the item was consistently more difficult for each response category for members of the F group than for comparable members of the R group, all three of the DIF procedures identified the item as having uniform DIF with similar power. For the smaller sample sizes of 500, the nominal form of the MH was less powerful than the ordinal MH extension. However, at the larger sample size of 2000, all of the procedures yielded high power estimates. The LDFA test for nonuniform DIF was nonsignificant, as it should have been for this DIF condition.

see Table 1 at end of report

For Condition 2, where item #20 was more discriminating for the R group for all response categories and the traditional nonuniform DIF was present, the LDFA test for nonuniform DIF showed moderate power for a sample size of 500 and higher power at the larger sample size. Two of the three uniform DIF tests (MH_{ord} and LDFA) showed very low power to detect this type of nonuniform DIF, as was to be expected. However, the MH_{nom} procedure showed moderate power (.30) in identifying this traditional nonuniform DIF at the larger sample size (2000). See Table 1.

For the third condition, where item #20 was somewhat consistently easier for the R group for the second category but more difficult for the third, the MH_{nom} procedure showed very high power to detect this type of nonuniform DIF even with the smaller sample size. The LDFA nonuniform DIF test had a low-to-moderate degree of power at the same sample size. Both the LDFA nonuniform test and the MH_{nom} demonstrated a high degree of estimated power for DIF identification at the larger sample size. Both the MH_{ord} and the uniform test of the LDFA procedure failed to identify this DIF situation in item# 20.

Table 2 gives estimates of average type I error for Items #1-#19 for each of the three different DIF procedures for the three DIF conditions. Recall that the nominal α level for these simulations was .0025. Table 2 shows that, with the exception of the LDFA nonuniform test for Condition 2, estimated type I error rates were within reasonable ranges of the nominal level for all procedures, for all sample sizes, and under all DIF conditions.

see Table 2 at end of report

Discussion and Conclusions

These simulations showed that the LDFA procedure was capable of identifying simulated DIF, both uniform and nonuniform, in polytomous items with a high degree of power. The procedure could also distinguish between uniform and nonuniform DIF. The only instance where the performance of the LDFA procedure was surpassed by another procedure was the condition simulated by Condition 3 when the sample sizes were fairly small. In this instance, the MH_{nom} procedure was much more sensitive to the directional change across response categories. However, with a larger sample size, the LDFA procedure also identified this type of nonuniform DIF accurately. The fact that the MH_{nom} procedure could not identify the type of nonuniform DIF simulated in Condition 2, even with fairly large samples of 2000 in each group, would suggest that it might not be the best procedure to use if the identification of such DIF is important. The MH_{ord} statistic was not accurate in identifying true DIF except in the uniform DIF situation. Even then, the LDFA approach was equally powerful in uncovering this type of DIF. Therefore, when fairly large sample sizes are available (i.e. $N > 500$), it is recommended that the LDFA procedure be used for DIF identification with polytomously scored test items.

References

Agresti, A. (1990). *Categorical Data Analysis.* New York: Wiley.

Holland, P. W., & Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure.* Princeton, NJ: Educational Testing Service, Research Report RR-86-31.

Hanson, B. (1992). *Comments on the Miller and Spray paper.* Internal memorandum, American College Testing.

Hauck, W. W. (1983). A note on confidence bands for the logistic response curve. *The American Statistician, 37*, 158-160.

Landis, J.R., Heyman, E.R., & Koch, G.G. (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests. *International Statistical Review, 46*, 237-254.

Mantel, N. (1963). Chi-square tests with one degree of freedom; Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Miller, T.R., & Spray, J.A. (1993) Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30,* 107-122.

Muraki, E. (1991). *A generalized partial credit model: Application of an EM algorithm.* Unpublished paper.

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

Table 1
Power Results, Item # 20

Condition	Sample Size	Procedure	Power	Average χ^2
1	500	MH_{nom} (3 df)	.350	13.016
		MH_{ord} (1 df)	.570	11.390
		LDFA *(uniform)* (1 df)	.600	12.136
		LDFA *(nonuniform)* (1 df)	.000	1.111
	2000	MH_{nom} (3 df)	1.000	40.172
		MH_{ord} (1 df)	1.000	38.177
		LDFA *(uniform)* (1 df)	1.000	38.715
		LDFA *(nonuniform)* (1 df)	.000	0.667
2	500	MH_{nom} (3 df)	.040	5.467
		MH_{ord} (1 df)	.020	1.869
		LDFA *(uniform)* (1 df)	.020	1.868
		LDFA *(nonuniform)* (1 df)	.440	9.686
	2000	MH_{nom} (3 df)	.300	11.693
		MH_{ord} (1 df)	.070	2.737
		LDFA *(uniform)* (1 df)	.060	2.692
		LDFA *(nonuniform)* (1 df)	1.000	31.196
3	500	MH_{nom} (3 df)	1.000	88.951
		MH_{ord} (1 df)	.000	1.393
		LDFA *(uniform)* (1 df)	.000	1.578
		LDFA *(nonuniform)* (1 df)	.370	8.857
	2000	MH_{nom} (3 df)	1.000	367.159
		MH_{ord} (1 df)	.000	1.945
		LDFA *(uniform)* (1 df)	.000	1.937
		LDFA *(nonuniform)* (1 df)	1.000	31.797

Table 2
Type I Error Results, Items #1-#19

Condition	Sample Size	Procedure	Error	Average χ^2
1	500	MH_{nom} (3 df)	.003	3.080
		MH_{ord} (1 df)	.005	1.056
		LDFA *(uniform)* (1 df)	.004	1.043
		LDFA *(nonuniform)* (1 df)	.004	1.106
	2000	MH_{nom} (3 df)	.004	3.146
		MH_{ord} (1 df)	.004	1.124
		LDFA *(uniform)* (1 df)	.004	1.121
		LDFA *(nonuniform)* (1 df)	.003	1.106
2	500	MH_{nom} (3 df)	.002	3.023
		MH_{ord} (1 df)	.002	1.106
		LDFA *(uniform)* (1 df)	.002	.993
		LDFA *(nonuniform)* (1 df)	.002	.970
	2000	MH_{nom} (3 df)	.002	2.977
		MH_{ord} (1 df)	.002	.996
		LDFA *(uniform)* (1 df)	.002	1.001
		LDFA *(nonuniform)* (1 df)	.007	1.237
3	500	MH_{nom} (3 df)	.005	3.045
		MH_{ord} (1 df)	.005	1.019
		LDFA *(uniform)* (1 df)	.005	1.012
		LDFA *(nonuniform)* (1 df)	.005	1.018
	2000	MH_{nom} (3 df)	.003	3.018
		MH_{ord} (1 df)	.002	.977
		LDFA *(uniform)* (1 df)	.002	.979
		LDFA *(nonuniform)* (1 df)	.001	1.009

Figure 1 ICCCs for Item #20, *Condition 1*

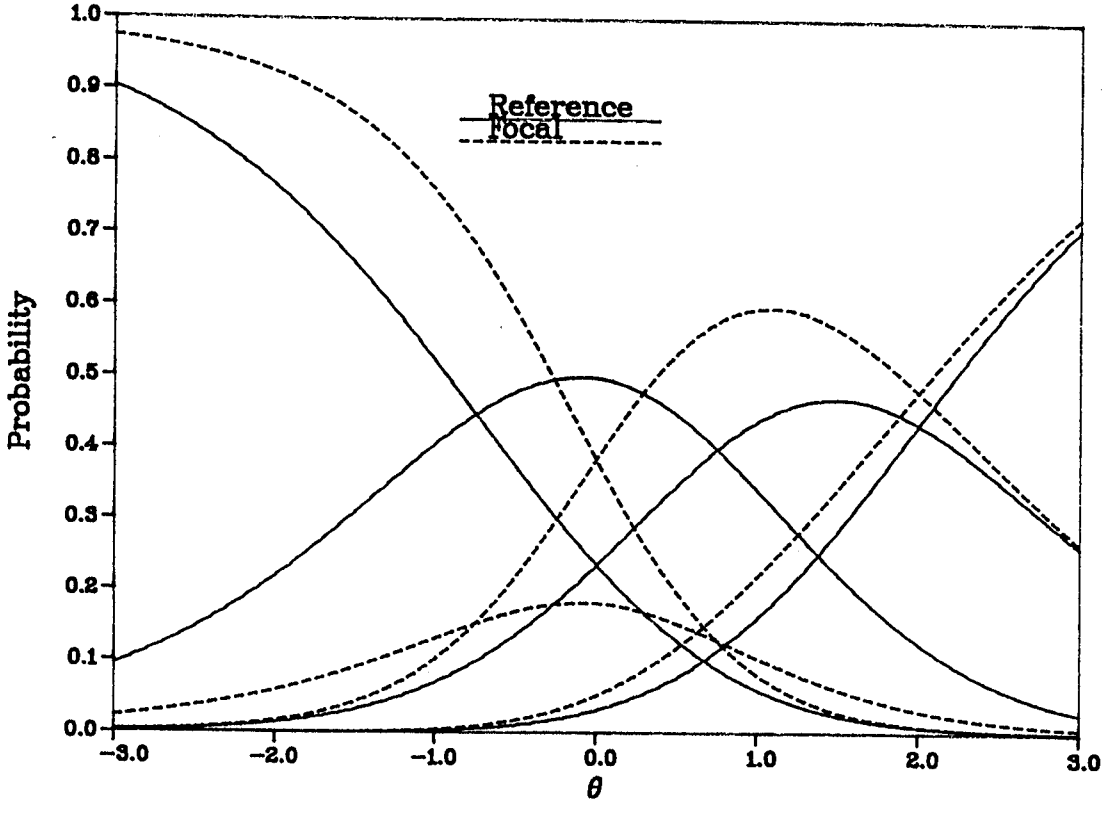

Figure 2 ICCCs for Item #20, *Condition 2*

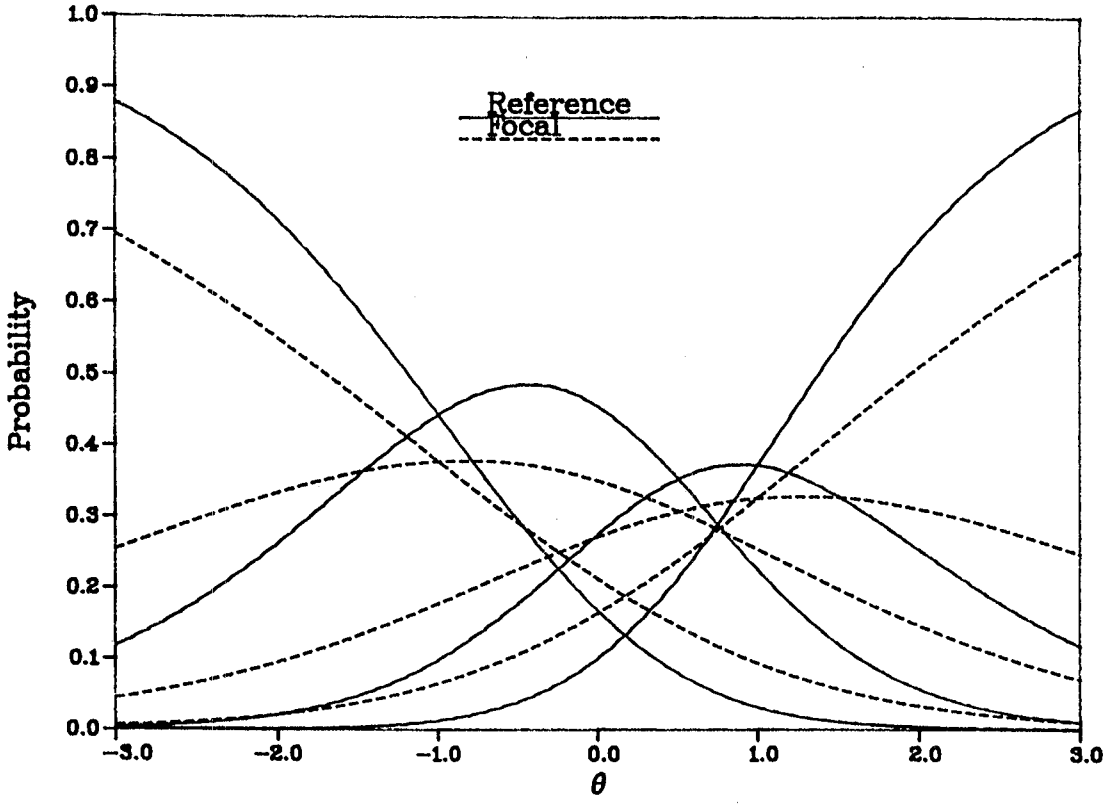
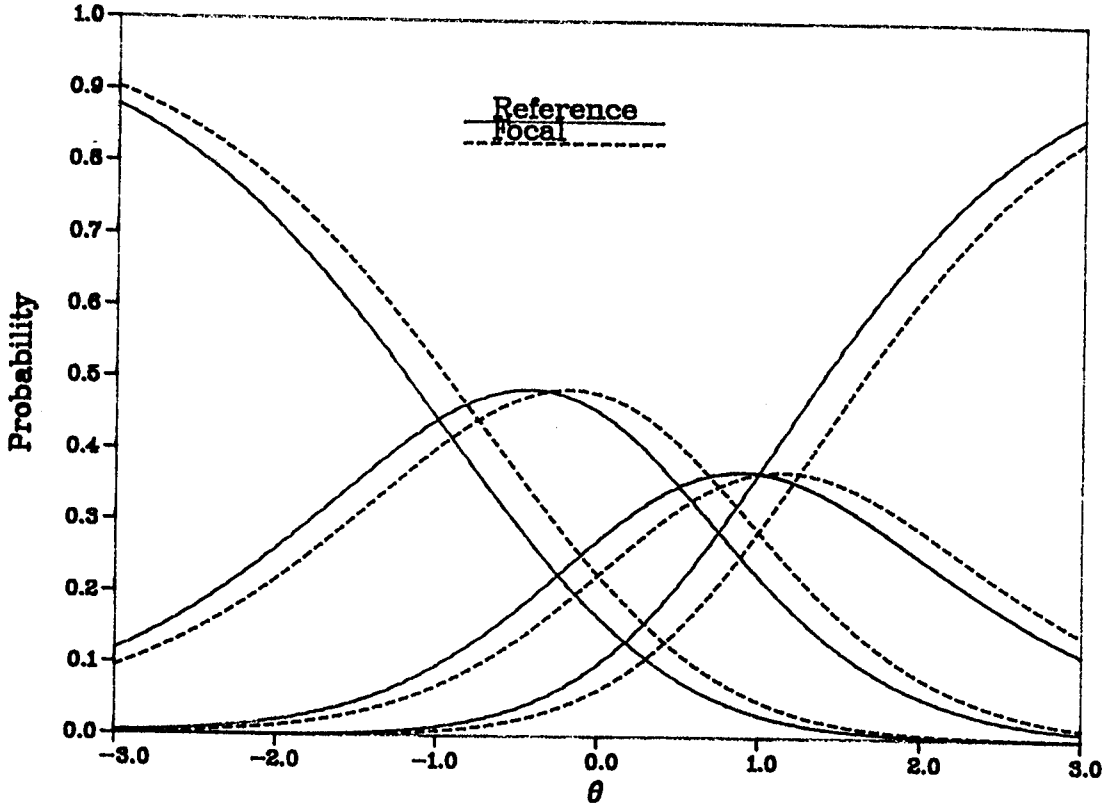
Figure 3 ICCCs for Item #20, *Condition 3*