ED 443 837 TM 031 482

AUTHOR	Kim, Seock-Ho; Cohen, Allan S.; DiStefano, Christine A.; Kim, Sooyeon
TITLE	An Investigation of the Likelihood Ratio Test for Detection of Differential Item Functioning under the Partial Credit Model.
PUB DATE	1998-04-14
NOTE	23p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Item Bias; Simulation; *Test Items
IDENTIFIERS	Item Bias Detection; *Likelihood Ratio Tests; *Partial Credit Model; Type I Errors

ABSTRACT

Type I error rates of the likelihood ratio test for the detection of differential item functioning (DIF) in the partial credit model were investigated using simulated data. The partial credit model with four ordered performance levels was used to generate data sets of a 30-item test for samples of 300 and 1,000 simulated examinees. Three different combinations of sample sizes of reference and focal group comparisons were simulated under two different ability matching conditions. One hundred replications of DIF detection comparisons were simulated for each of six conditions. Type I error rates of the likelihood ratio for all six conditions were with theoretically expected values at each of the nominal alpha levels considered. (Contains 25 references.) (Author/SLD)

An Investigation of the Likelihood Ratio Test for Detection of Differential Item Functioning Under the Partial Credit Model

Seock-Ho Kim
The University of Georgia

Allan S. Cohen
University of Wisconsin–Madison

Christine A. DiStefano
Sooyeon Kim
The University of Georgia

April 14, 1998

Running Head: LIKELIHOOD RATIO TEST UNDER
THE PARTIAL CREDIT MODEL

Paper presented at the annual meeting of the American Educational
Research Association, San Diego, California.

An Investigation of the Likelihood Ratio Test for Detection of Differential Item Functioning Under the Partial Credit Model

Abstract

Type I error rates of the likelihood ratio test for the detection of differential item functioning (DIF) in the partial credit model were investigated using simulated data. The partial credit model with four ordered performance levels was used to generate data sets of a 30-item test for samples of 300 and 1,000 simulated examinees. Three different combinations of sample sizes of reference and focal group comparisons were simulated under two different ability matching conditions. 100 replications of DIF detection comparisons were simulated for each of the six conditions. Type I error rates of the likelihood ratio test for all six conditions were within theoretically expected values at each of the nominal alpha levels considered.

Index terms: differential item functioning, item response theory, likelihood ratio test, partial credit model, Type I error.

Introduction

Under item response theory (IRT), a dichotomous item is said to be functioning differentially, when the probability of a correct response to the item is different for examinees at the same ability level but from different groups (Pine, 1977). For polytomous IRT models, differential item functioning (DIF) is present, when the item true score functions in different groups are not equal (Cohen, Kim, & Baker, 1993). The existence of such items on a test is a threat to validity and may seriously interfere with efforts to equate tests.

The likelihood ratio (LR) test (Neyman & Pearson, 1928) has been proposed by Thissen, Steinberg, and Gerrard (1986) and by Thissen, Steinberg, and Wainer (1988, 1993) for detection of DIF. This use of the LR test evaluates the equality of item parameters estimated in the different groups. Kim and Cohen (1995) compared this LR test with Lord's (1980) chi-square test, and also with Raju's area measures (1988, 1990) for the dichotomous model and found them all to provide comparable results. Cohen, Kim, and Wollack (1996) subsequently reported Type I error rates of the LR test for DIF under the two- and three-parameter IRT models to be within expected limits at the nominal alpha levels considered.

Relatively few studies of Type I error rates or power for the LR test for DIF have been done, however, with any polytomous models. Ankenmann, Witt, and Dunbar (1996) compared Type I error and power for the LR test and the Mantel (1963) test for DIF detection on a test composed of both dichotomous and graded response items. Type I error rates and power for the LR test were obtained for a single studied graded response item on each test data set under the different sample size and ability conditions simulated. The LR test was found to yield better power and Type I error control than the Mantel test (Ankenmann et al., 1996). Kim and Cohen (in press) also reported that Type I errors were controlled at the nominal level of significance for the LR test for detection of DIF under the graded response model.

Results for the graded response model are useful but it is not clear that the results from Ankenmann et al. (1996) and Kim and Cohen (in press) generalize directly to other polytomous IRT models such as the partial credit model (Masters, 1982). Although the partial credit model may provide a useful alternative to the graded response model for the analysis of performance type assessments (Fitzpatrick, Link, Yen, Burket, Ito, & Sykes, 1996; Zwick, Donoghue, & Grima, 1993), and although the partial credit model has been employed in investigations of DIF for polytomously scored items (Chang, Mazzeo, & Roussos, 1996;

Zwick et al., 1993), no results have yet been reported on the applicability of the LR test for this model. In this paper, we investigate the Type I error control of the LR test of DIF for the partial credit model, under the same conditions used by Kim and Cohen (in press) for the graded response model.

It is important to recognize in this context that investigation of the power of a statistical test is meaningless if adequate control of the Type I error rate is not maintained. As previous research does not provide sufficient information about the Type I error control for the LR test for DIF for the partial credit model, the present study was designed to examine the Type I error rates of the LR test under the partial credit model for a variety of underlying item parameters, sample sizes and ability conditions.

Item True Score Functions

The item response function (IRF) for a dichotomous item is the same as the item true score function that describes the functional relationship between the probability of a correct response to an item and examinee trait level, θ. The item true score function for a polytomous item describes the relationship between the expected value of the item score and θ. For both dichotomous and polytomous IRT models, an item functions differentially, if the item true score functions obtained from different groups of examinees are different. Item true score functions will be identical, however, if the sets of item parameters estimated in different groups are equal.

For polytomous IRT models, there are several approaches to testing the equality of sets of items. One approach is to compare item parameters estimated in different groups (e.g., Cohen, Kim, & Baker, 1993). A second approach is to compare areas between item true score functions estimated in different groups (e.g., Cohen, Kim, & Baker, 1993; Flowers, Oshima, & Raju, 1995). A third approach is to compare likelihood functions, using a LR test. Thissen et al. (1988) suggest that this third approach is preferable, as the first and second approaches may require estimates of variances and covariances of the item parameters, which may be not be accurately estimated by current algorithms.

Establishing a Common Metric

DIF studies under IRT require that estimates of item parameters obtained in different groups be placed on a common metric before comparisons are made (see Stocking & Lord, 1983). Such transformations or linking of metrics are unnecessary for the LR test of DIF using the computer program MULTILOG (Thissen, 1991), however, because item parameters are estimated simultaneously in a single data set consisting of both the reference and focal groups combined. The common metric for conducting the LR test for DIF is obtained through the common or anchor set of items rather than by linking. In the LR test, the likelihood from a compact model, in which no group differences are assumed to be present, is compared to that from an augmented model, in which one or more items are examined for possible DIF. The metric of the compact and augmented models are dependent on the anchor items. The assumption in this approach is that there are no DIF items among the common items in the compact model.

The comparison between a compact model and an augmented model for the LR test requires two separate calibration runs for obtaining the likelihoods. Thissen et al. (1993) recommended the use of the Mantel Haenszel (MH) χ^2 to identifying a set of common items which are non-DIF for purposes of establishing a common metric for dichotomous models. This approach can be useful but is also suspect as the MH test can not detect non-uniform DIF. Kim and Cohen (1995) recommended an iterative purification method for the likelihood ratio test. Although the iterative purification method is quite labor intensive, it is theoretically more consistent with the likelihood ratio test. For polytomous IRT models one can apply either Mantel's (1963) test or some type of method such as suggested by Kim and Cohen (1995).

Method

The simulation and DIF detection procedures used in this study are similar to those used by Kim and Cohen (in press) for the graded response model.

Data Generation

In the usual DIF study, there are two groups of examinees, the reference group and the focal group. The reference group is the base group against which the parameters estimated in the focal group are compared. For the reference and focal groups in this study, two sample sizes were used to simulate small sample ($N = 300$) and large sample ($N = 1,000$) conditions. Three different sample size combinations of reference and focal groups were simulated: (1) a reference group with 300 examinees and a focal group with 300 examinees (R300/F300), (2) a reference group with 1,000 examinees and a focal group with 1,000 examinees (R1000/F1000), and (3) a reference group with 1,000 examinees and a focal group with 300 examinees (R1000/F300). The large and small sample size comparisons were selected based on previous recovery study results for the partial credit model by Choi, Cook, and Dodd (1997) which indicated that at least 250 examinees were needed to achieve an adequate calibration of item step parameters.

The computer program RESGEN (Muraki, 1996) was used to generate data sets for the partial credit model (Masters, 1982; Masters & Wright, 1997) for a 30-item test with four ordered performance levels. The category response function $P_{jx}(\theta)$ is the probability of response $X_j = x$ to item j as a function of θ, and can be defined as

$$\text{Prob}\left[X_j = x|\theta, \delta_{j1}, \ldots, \delta_{m_j}\right] = P_{jx}(\theta) = \frac{\exp\left[\sum\limits_{k=0}^{x}(\theta - \delta_{jk})\right]}{\sum\limits_{h=0}^{m_j}\exp\left[\sum\limits_{k=0}^{h}(\theta - \delta_{jk})\right]}, \tag{1}$$

where $x = 0(1)m_j$, $\theta - \delta_{j0} = 0$, and $\sum\limits_{x=0}^{m_j} P_{jx}(\theta) = 1$. For each item with $m_j + 1$ ordered performance levels with scores of $0, 1, \ldots, m_j$, there exist m_j item parameters, that is, $\delta_{j1}, \delta_{j2}, \ldots, \delta_{jm_j}$. Each item parameter δ_{jk} corresponds to the point on the θ scale where the probability of category k is the same as the probability of category $k - 1$. Equivalently, at δ_{jk}, there will be the same probability of observing the item scores x and $x - 1$. As an example, for an item j with four ordered performance levels (i.e., $x = 0, 1, 2, 3$), the item parameters $\delta_{j1}, \delta_{j2}, \delta_{j3}$ are the points on the θ scale where the respective item category response functions of 0 and 1, of 1 and 2, and of 2 and 3 intersect.

The expected score of item j give θ is defined as

$$E(X_j|\theta) = \sum_{x=0}^{m_j} x P_{jx}(\theta). \tag{2}$$

In a DIF study, we test the null hypothesis that

$$E_{\text{R}}(X_j|\theta) = E_{\text{F}}(x_j|\theta), \tag{3}$$

where R and F designate the reference group and the focal group, respectively.

Item parameter values with four ordered performance levels were used to generate 100 replications of the partial credit model data based on the item parameters reported by Koch and Dodd (1989). The generating item parameters used in this study are given in Table 1.

Insert Table 1 about here

For each of the three sample sizes, two different ability matching conditions were simulated: (1) a matched condition in which both the reference and focal groups of examinees had the same underlying ability distribution [$\theta \sim$ N(0,1)], and (2) an unmatched condition in which the reference group had a higher underlying ability distribution [$\theta \sim$ N(0,1)] than that of the focal group [$\theta \sim$ N($-$1,1)]. 100 replications were simulated for each of the six combinations of three sample sizes by two ability matching conditions.

The Likelihood Ratio Test

Item parameter estimates for each pair of reference and focal groups were obtained using the default options available from the marginal maximum likelihood estimation algorithm for the partial credit model as implemented in the computer program MULTILOG (Thissen, 1991). We adopt the convention for DIF detection in this study that a studied item in the focal group is compared with the item with the same generating parameters in the reference group.

The LR test for DIF described by Thissen et al. (1988, 1993) compares two different models—a compact model and an augmented model. The LR statistic, G^2, is the difference between the values of -2 times the log likelihood for the compact model (L_{C}) and -2 times the log likelihood for the augmented model (L_{A}). Values of the quantity -2 times the log likelihood can be obtained from the output of MULTILOG and are based on the results over the entire data set following marginal maximum likelihood estimation. G^2 can be written as

$$G^2 = -2\log L_{\mathrm{C}} - (-2\log L_{\mathrm{A}}) = -2\log L_{\mathrm{C}} + 2\log L_{\mathrm{A}} \qquad (4)$$

and is distributed as a χ^2 under the null hypothesis with degrees of freedom equal to the difference in the number of parameters estimated in the compact and augmented models. In this study, one item was tested at a time. This meant that each G^2 was distributed as a χ^2 with 3 degrees of freedom.

In the compact model, the item parameters are assumed to be the same for both the reference and focal groups. Options in MULTILOG permits equality constraints to be placed on item parameters. In this study, the parameter estimates for all 30 items in the compact model were constrained to be equal in both the reference and focal groups. Similarly, for the augmented model, all items except the studied item were constrained to be equal in the reference and focal groups. Constrained items in the augmented model form the common or anchor set of items. For DIF comparison simulated in this study, only the item parameters for the studied item were unconstrained, that is, the item parameters for the studied item were allowed to assume different values in the reference and focal groups. For an augmented model in which Item 1 was the studied item, for example, the item parameter estimates for Item 1 were unconstrained in both the reference and focal groups. In this augmented model, Items 2–30 formed the anchor set and consequently were each constrained to have the same parameter estimates in both groups. As indicated above, DIF comparisons in this study were simulated to study only a single item at a time.

Error Rates

Error rates for the LR test were obtained by comparing the number of significant G^2s to the total number of augmented model calibration runs conducted for a given sample size and ability matching condition. For a single test, 31 separate calibration runs were required to estimate the necessary likelihood statistics, one run to estimate the likelihood for the compact model and 30 runs for each of the augmented models (i.e., one augmented model for each of the 30 items). For the 100 pairs of reference and focal groups in a sample size by ability matching condition, 3,100 separate calibration runs were required. A total of 18,600 MULTILOG calibration runs were required for all six sample sizes by ability matching conditions.

Results

False Positive Errors

The numbers of significant G^2s for each item at $\alpha = .05$ are given in Table 1. These data illustrate the general pattern of results obtained in the conditions simulated in this study. For Item 1, in the R300/F300 condition, for example, 3 significant G^2s were obtained for the matched ability condition and 8 for the unmatched ability condition. Since 100 replications were generated, the expected number of significant G^2s due to chance for a single item would be 5 at a nominal alpha of .05. For this same sample size condition, there were a total of 150 significant G^2s obtained across all 30 items for the matched ability condition and 163 for the unmatched ability condition. A similar pattern of results was found at all other α levels examined.

The numbers of significant G^2s for the different alpha levels for all sample size and ability matching conditions are given in Table 2. The R300/F300 sample size at $\alpha = .05$, for example, yielded 150 significant G^2s for the matched ability condition and 163 for the unmatched ability condition. The expected number of significant G^2s due to chance over the 100 replications at a nominal alpha level of .05 would be 150 (i.e., 30 items times 100 replications time .05).

Insert Table 2 about here

The bottom row of Table 2 contains the expected number of significant G^2s for the alpha levels considered in this study. The observed numbers of significant G^2s at each alpha level appear to be very close to the theoretically expected values for all the sample size by ability matching conditions.

Type I Error Rates

Type I error rates are given in Table 3 as the proportion of significant G^2s at each alpha level over all replications. Error rates for the three sample sizes are illustrated in Figures 1a, and 1b, for the matched ability condition and the unmatched ability condition, respectively.

Insert Table 3 and Figures 1a and 1b about here

Type I error rates appeared to be only slightly elevated for the matched ability conditions, for R300/F300 at the .0005 and .1 nominal alpha levels and for R1000/F300 at the .0005, .001, .005, and .01 levels. For the unmatched ability condition, R300/F300 yielded similarly very slightly inflated Type I error rates for the .1 nominal alpha level. The R1000/F1000 unmatched ability condition yielded slightly lower error rates than expected at all alpha levels as did the R1000/F300 DIF comparisons for the unmatched condition.

Insert Figures 2a, 2b, and 2c about here

Results of error rates for the two ability matching conditions for R300/F300, R1000/F1000, and R1000/F300 are illustrated in Figures 2a, 2b and 2c, respectively. For both R300/F300 and R1000/F300 comparisons, as can be seen in these figures, Type I error rates were a bit closer to the theoretically expected values in the matched ability condition at the .05

nominal alpha level. For R1000/F1000 in the unmatched ability condition, the Type I error rate was slightly closer to the theoretically expected value at the .05 nominal alpha level. It is very important to note, however, that all deviations from expectations were very small. Most of the Type I error rates, in fact, were quite close to the theoretically expected values in all conditions simulated.

Relationships Among Generating Parameters and Significant G^2s

Spearman rank-order correlations are used to describe the relationships among the generating parameters and the number of significant G^2s at the nominal alpha level of .05 (see Table 4). Correlations among item parameter estimates were all positive, and no consistent pattern of relationships was observed between generating item parameters and the numbers of significant G^2s.

Insert Table 4 about here

Summary and Discussion

Type I error rates for G^2 for the partial credit model were very close to those expected for all sample sizes and alpha levels considered in both ability matching conditions. Results for the small sample size condition, however, did differ slightly from expected values at the .1 nominal alpha level in both ability matching conditions. Even so, Type I error rates for the small sample condition R300/F300 were quite close to the theoretically expected values. These results are in agreement with similar results from Kim and Cohen (in press) for the LR test for DIF for the graded response model.

Generally, in most DIF studies, it is desirable to detect all items that function differentially so as to be able to remove them from the equating process or to study them further for subsequent bias analysis. This is typically accomplished by setting the nominal alpha level high, for example, at .05 or .1 under the assumption that it is more preferable to falsely identify an item as functioning differentially than it is to miss a true DIF item. At such alpha levels, the LR test was found to provide Type I error control very close to the nominal level for the sample sizes and ability matching conditions simulated. There is also a concern in some DIF studies, however, regarding for the power of the DIF statistic. Power is low when true DIF items are not detected. Studies examining the power of polytomous models such as the partial credit model are clearly needed.

In this study, the underlying ability distribution for the reference group was set to be N(0,1). The reason for this choice was to match the distribution of ability with that of the item step parameters. There were two different underlying distributions for the focal group: N(0,1) for the matched condition and N(−1,1) for the matched condition. No systematic differences in the Type I error rates were observed between the matched and unmatched ability conditions.

The LR test of DIF using a the computer program MULTILOG permits concurrent calibration, and so does not require any metric transformation. The resulting Type I error rates, therefore, do not contain errors due to linking. However, the influence of DIF items in the anchor set is still a potential problem. Such items are likely to affect Type I error control and, consequently, the power of the LR test statistic. Methods for construction of the anchor sets of items and for scale purification with the likelihood ratio test need to be studied.

References

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1996, April). *An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33,* 333–353.

Choi, S. W., Cook, K. F., & Dodd, B. G. (1997). Parameter recovery for the partial credit model using MULTILOG. *Journal of Outcome Measurement, 1,* 114–142.

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17,* 335–350.

Cohen, A. S., Kim, S.-H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20,* 15–26.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1995, April). *A Monte Carlo assessment of DFIT with polytomously scored unidimensional tests.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Fitzpatrick, A. R., Link, V. B., Yen, W. M., Burket, G. R., Ito, K., & Sykes, R. C. (1996). Scaling performance assessment: A comparison of one-parameter and two-parameter partial credit models. *Journal of Educational Measurement, 33,* 291–314.

Kim, S.-H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ration test in detection of differential item functioning. *Applied Measurement in Education, 8,* 291–312.

Kim, S.-H., & Cohen, A. S. (in press). An investigation of the likelihood ratio test for detection of differential item functioning under the graded response model. *Applied Psychological Measurement.*

Koch, W. R., & Dodd, B. G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education, 2,* 335–357.

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

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.

Masters, G. N. (1982). A Rasch model for partial credit scoring, *Psychometrika, 47,* 149–174.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). New York: Springer.

Muraki, E. (1996). *RESGEN: Item parameter generator* [Computer program]. Princeton, NJ: Educational Testing Service.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I and Part II. *Biometrika, 20A,* 174–240, 263–294.

Pine, S. M. (1977). Application of item characteristic curve theory to the problem of test bias. In D. J. Weiss (Ed.), *Application of computerized adaptive testing: Proceedings of a symposium presented at the 18th annual convention of the Military Testing Association* (Research Rep. No. 77-1, pp. 37–43). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

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,* 197–207.

Stocking, M., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 207–210.

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory* [Computer program]. Chicago: Scientific Software.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118–128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Erlbaum.

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. 67–113). Hillsdale, NJ: Erlbaum

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

Table 1

Generating Item Parameters and Number of Significant G^2s at $\alpha = .05$ for Sample Size and Ability Matching Conditions

	Parameters			R300/F300		R1000/F1000		R1000/F300	
Item	δ_{j1}	δ_{j2}	δ_{j3}	Matched	Unmatched	Matched	Unmatched	Matched	Unmatched
1	−1.00	0.00	1.00	3	8	5	5	2	3
2	−1.35	0.00	1.35	2	3	6	3	7	4
3	−1.25	0.00	1.25	3	3	5	10	4	4
4	.00	−1.25	1.25	7	7	5	3	6	6
5	−1.25	1.25	0.00	10	6	4	4	1	8
6	−1.00	0.00	1.00	5	7	4	3	6	4
7	−1.35	0.00	1.35	6	4	6	2	8	7
8	−1.25	0.00	1.25	2	8	5	3	7	2
9	0.00	−1.25	1.25	6	7	5	7	5	6
10	−1.25	1.25	0.00	6	6	3	4	6	6
11	0.50	1.50	2.50	5	6	3	4	6	3
12	0.50	1.75	2.50	5	6	4	5	3	5
13	0.70	2.00	2.70	3	2	6	10	6	3
14	0.80	1.90	2.50	4	4	5	7	4	3
15	0.80	1.40	2.50	5	5	2	5	5	9
16	0.50	0.90	2.50	10	10	3	3	2	3
17	1.75	0.50	2.50	5	2	7	2	4	5
18	0.50	2.50	1.75	4	7	8	6	2	6
19	1.40	0.80	2.50	3	4	3	8	1	3
20	0.80	2.50	1.40	5	4	3	3	4	7
21	−2.50	−1.50	−0.50	6	8	4	4	4	1
22	−2.50	−1.75	−0.50	3	4	8	4	10	5
23	−2.70	−2.00	−0.70	1	6	5	3	5	5
24	−2.50	−1.90	−0.80	3	2	4	8	6	6
25	−2.50	−1.40	−0.80	8	5	4	4	4	3
26	−2.50	−0.90	−0.50	8	7	1	5	2	3
27	−2.50	−0.50	−1.75	3	4	6	5	6	2
28	−1.75	−2.50	−0.50	6	9	3	4	5	8
29	−2.50	−0.80	−1.40	6	2	3	5	4	4
30	−1.40	−2.50	−0.80	7	7	5	3	8	4
Total				150	163	135	142	143	138

Table 2

Number of Significant G^2s for Sample Size and Ability Matching Conditions at α Levels From .0005 to .1

		α Level					
Sample Size	Ability	.0005	.001	.005	.01	.05	.1
R300/F300	Matched	2	2	13	29	150	327
R300/F300	Unmatched	2	3	15	28	163	324
R1000/F1000	Matched	1	2	13	36	135	295
R1000/F1000	Unmatched	0	1	12	28	142	289
R1000/F300	Matched	2	4	19	37	143	294
R1000/F300	Unmatched	0	1	9	23	138	297
Expected Value		1.5	3	15	30	150	300

Table 3

Proportion of Significant G^2s for Sample Size and Ability Matching Conditions at α Levels From .0005 to .1

		α Level					
Sample Size	Ability	.00050	.00100	.00500	.01000	.05000	.10000
R300/F300	Matched	.00067	.00067	.00433	.00967	.05000	.10900
R300/F300	Unmatched	.00067	.00100	.00500	.00933	.05433	.10800
R1000/F1000	Matched	.00033	.00067	.00433	.01200	.04500	.09833
R1000/F1000	Unmatched	.00000	.00033	.00400	.00933	.04733	.09633
R1000/F300	Matched	.00067	.00133	.00633	.01233	.04767	.09800
R1000/F300	Unmatched	.00000	.00033	.00300	.00767	.04600	.09900

Table 4

Spearman ρs Among Generating Item Parameters and Number of Significant G^2s at $\alpha = .05$

		Parameters			R300/F300		R1000/F1000		R1000/F300	
Parameter or Condition		δ_{j1}	δ_{j2}	δ_{j3}	Matched	Unmatched	Matched	Unmatched	Matched	Unmatched
δ_{j1}		1.000	.784	.897	−.041	−.069	−.042	.102	−.297	.119
δ_{j2}			1.000	.784	−.100	−.189	−.064	.177	−.339	.046
δ_{j3}				1.000	−.143	−.098	.040	.067	−.199	.044
R300/F300	Matched				1.000	.378	−.449	−.244	−.249	.211
	Unmatched					1.000	−.244	−.250	−.161	−.097
R1000/F1000	Matched						1.000	−.094	.371	−.056
	Unmatched							1.000	−.349	−.145
R1300/F1300	Matched								1.000	.077
	Unmatched									1.000

Figure Captions

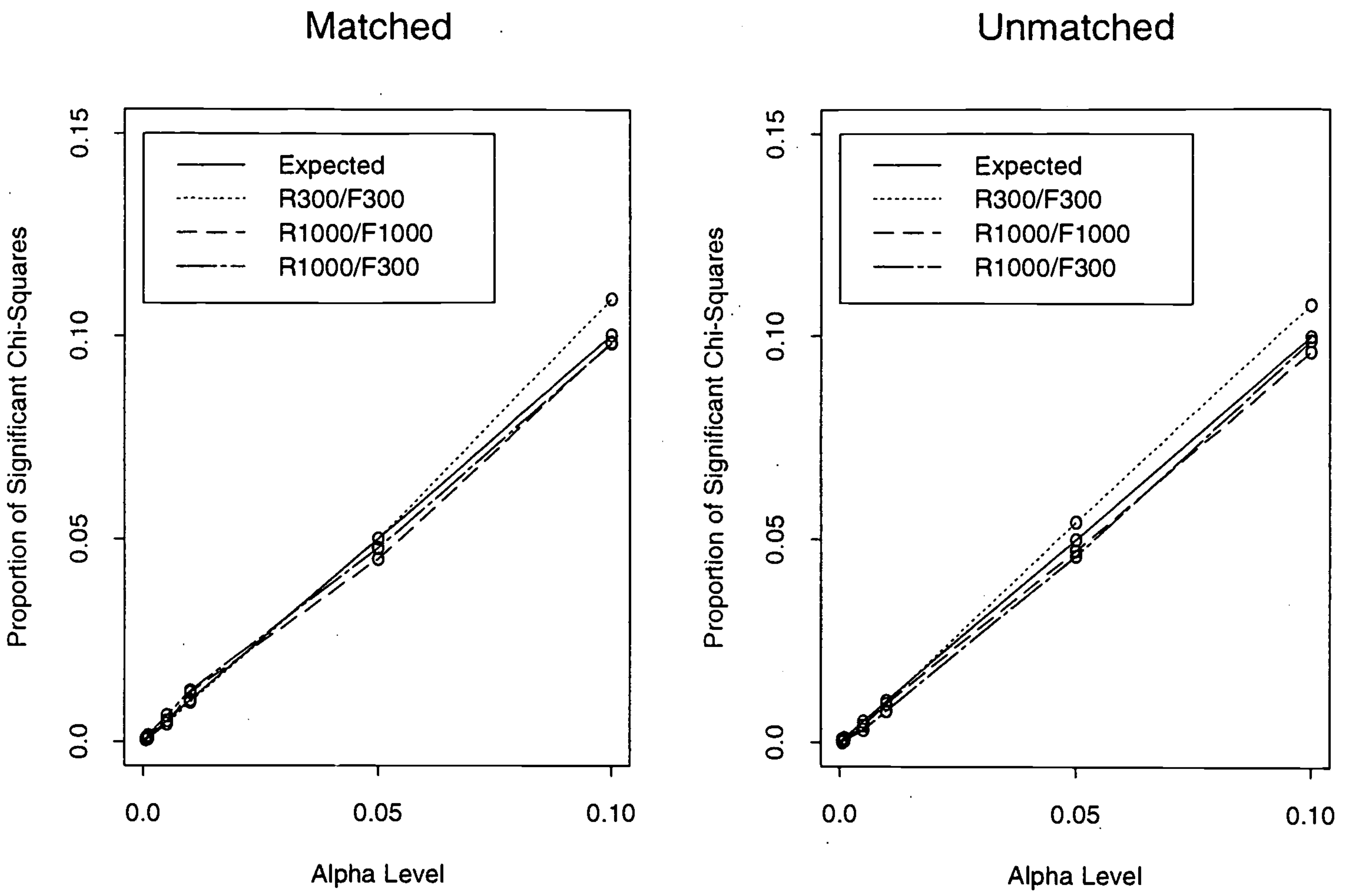
Figure 1a. Proportion of Significant G^2s for the Matched Ability

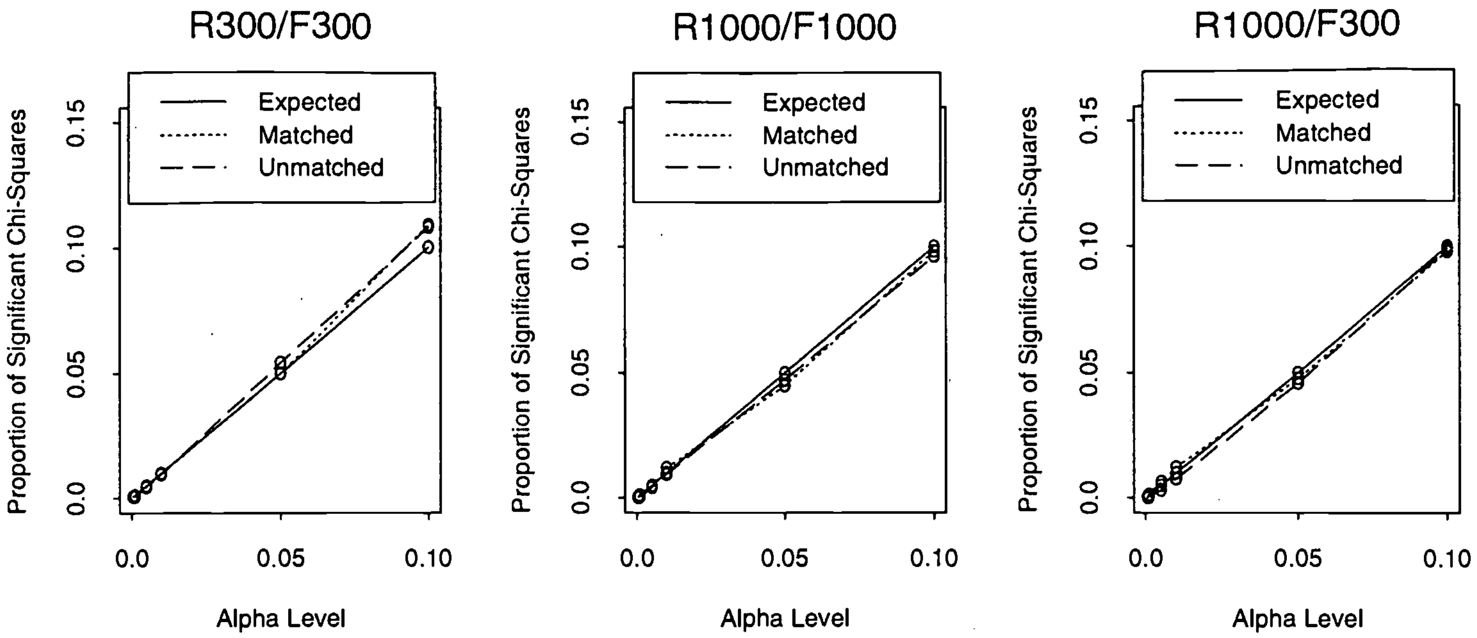
Figure 1b. Proportion of Significant G^2s for the Unmatched Ability

Figure 2a. Proportion of Significant G^2s for the R300/F300

Figure 2b. Proportion of Significant G^2s for the R1000/F1000

Figure 2c. Proportion of Significant G^2s for the R1000/F300


Matched
Unmatched
Expected
R300/F300
R1000/F1000
R1000/F300
Proportion of Significant Chi-Squares
0.0
0.05
0.10
0.15
Alpha Level
0.0
0.05
0.10


R300/F300
R1000/F1000
R1000/F300
Expected
Matched
Unmatched
Proportion of Significant Chi-Squares
0.0
0.05
0.10
0.15
Alpha Level
0.0
0.05
0.10

Acknowledgments

The authors express their gratitude to Eiji Muraki for making his computer program RESGEN available for this study and to David Thissen for his help in setting up MULTILOG input files.

Author's Address

Send all correspondence to Seock-Ho Kim, The University of Georgia, 325 Aderhold Hall, Athens, GA 30602, U.S.A. Internet: skim@coe.uga.edu

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: An Investigation of the Likelihood Ratio Test for Detection of Differential Item Functioning Under the Partial Credit Model

Author(s): Seock-Ho Kim, Allan S. Cohen, Christine A. DiStefano, & Sooyeon Kim

Corporate Source: The University of Georgia AERA

Publication Date: April, 1998

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	Level 2A	Level 2B
☑	□	□
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproductio n by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: SeockHo Kim

Printed Name/Position/Title: Seock-Ho Kim, Assistant Professor

Organization/Address: The University of Georgia
325 Aderhold Hall
Athens, GA 30602

Telephone: (706) 542-4224

FAX: (706) 542-4240

E-Mail Address: skim@coe.uga.edu

Date: 9/3/98

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

The Catholic University of America
ERIC Clearinghouse on Assessment and Evaluation
210 O'Boyle Hall
Washington, DC 20064
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com