ED 374 153 TM 022 082

AUTHOR	Pang, Xiao L.; And Others
TITLE	Performance of Mantel-Haenszel and Logistic Regression DIF Procedures over Replications Using Real Data.
PUB DATE	Apr 94
NOTE	18p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	College Entrance Examinations; Foreign Countries; Identification; *Item Bias; Mathematics Tests; Performance; *Regression (Statistics); *Sample Size; Scores; *Sex Differences; Test Items; White Students
IDENTIFIERS	ACT Assessment; Criterion Variables; *Logistic Regression; *Mantel Haenszel Procedure; Power (Statistics); Research Replication

ABSTRACT

The function of Mantel-Haenszel (MH) and logistic regression (LR) statistics with real data in detecting gender-based differentially functioning items (DIF) was investigated when sample size and criterion variable varied. The data base consisted of the item responses of a population of 183,356 Caucasians to the Math test of the ACT Assessment (Form 39B) in 1989. Using four sample sizes (1000/1000, 500/500, 250/250, and 100/100) and two criterion scores (total test and subtest), 30 replications were carried out for each combination of sample size and criterion. The results indicated that the power of the two procedures was affected by sample size and the choice of criterion, when items with different magnitude of DIF were assessed. While displaying a high agreement in detection rate for each condition, the extent of DIF decreased for 10 out of 12 DIF items when subtest score was used as the criterion. (Contains 10 references and 2 tables.) (Author)

PERFORMANCE OF MANTEL-HAENSZEL AND LOGISTIC REGRESSION DIF PROCEDURES OVER REPLICATIONS USING REAL DATA

XIAO L. PANG, FANG TIAN AND MARVIN W. BOSS

UNIVERSITY OF OTTAWA

RUNNING HEAD: SAMPLE SIZE AND CRITERION EFFECT ON DIF

Abstract

The function of Mantel-Haenszel (MH) and logistic regression (LR) statistics with real data in detecting gender-based differentially functioning items (DIF) was investigated when sample size and the criterion variable varied. The data base used was the item responses of a population of 183,356 Caucasians to the Math test of the ACT Assessment (Form 39B) in 1989. Using four sample sizes (1000/1000, 500/500, 250/250, and 100/100) and two criterion scores (total test and subtest), thirty replications were carried out for each combination of sample size and criterion. The results indicated that the power of the two procedures was affected by sample size and the choice of criterion, when items with different magnitude of DIF were assessed. While displaying a high agreement in detection rate for each condition, the extent of DIF decreased for 10 out of 12 DIF items when subtest score was used as the criterion.

Performance of Mantel-Haenszel and Logistic Regression DIF Procedures Over Replications Using Real Data

Differential item functioning (DIF), a major threat to test validity, has been an important issue in educational measurement. Many methods, including item response theory (IRT) techniques, have been developed for detecting DIF. Although IRT-based methods are theoretically preferred, several disadvantages, such as sample size requirement, strict IRT assumptions, unstable parameter estimates, and lack of statistical test, make it difficult to implement practically.

As an alternative to IRT-based approaches, the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1986) has become one of the most popular techniques in identifying DIF because of its computational simplicity, ease of implementation, and associated test of statistical significance. In addition, it provides information on the magnitude as well as the direction of DIF. However, the MH procedure also has its limitations. Since it is specially designed to detect uniform DIF, the MH procedure may be less powerful in detecting nonuniform DIF. A logistic regression (LR) procedure (Swaminathan & Rogers, 1990) has aroused more and more interest among researchers as a result of its use in identifying not only uniform DIF but also nonuniform DIF.

The findings from most studies assessing the performance of either procedure suggest that sample size affects the power of the MH and LR procedures. For example, Swaminathan and Rogers (1990) found in their simulated study that, with uniform DIF, the MH and LR procedures identified from 75% of DIF items with small sample size and short test length to 100% with large sample size and long test length. With nonuniform DIF, the LR procedure identified 50% of DIF items with small sample size and short test length and 75% with large sample size and long test length. The MH procedure completely failed to identify nonuniform DIF under any conditions.

In their extensive comparison of the performance of the MH and LR procedures, Rogers and Swaminathan (1993) showed that sample size, size of DIF, and type of DIF affected the power of the two DIF indices. With uniform DIF increasing sample size substantially increased the detection rates for both procedures. The highest detection rates were found for the items with moderate difficulty and high discrimination and items with larger DIF effects. With nonuniform DIF, a similar pattern of effect of sample size, amount of DIF, and type of DIF was observed. The MH procedure was almost completely unable to detect strictly (disordinal) nonuniform DIF. For items of low difficulty (ordinal nonuniform DIF), the MH detection rate was also substantially lower than the LR detection rate.

According to Ackerman (1992), multidimensionality of a test may contribute to the occurrence of DIF. This position is closely related to the issue of choice of criterion variable. So far, studies concerning choice of criterion variable have yielded inconsistent results. Hambleton, Bollwark, and Rogers (1990) found that the MH results were quite similar for internal and external criteria. Tian, Pang, and Boss (1994) showed that choice of total test and subtest as the matching criterion did not affect the detection rate for either procedure. However, Clauser, Mazor, and Hambleton (1991) showed that the choice of criterion, total test score or subtest score, had a substantial influence on the classification of items using the MH procedure. Twenty-two items showed DIF with total test. However, seven out of 22 items ceased to show DIF and 11 new DIF items were found when subtests were used.

There is limited information available regarding the agreement between these two procedures when real data are used. In addition, we do not know what minimum sample size is necessary to detect DIF and how the choice of different criterion scores affects the performance of these indices using a large data base.

The purpose of this study is to examine the consistency of the performance of the MH and LR statistics across different sample sizes using total test score and subtest score as criterion over replications. Three research questions are addressed in this study:

1) How large a sample size is needed to detect DIF?

2) Is the identification of DIF by the MH and LR procedures affected by choice of total test score or subtest score?

3) How consistently do the MH and LR procedures agree?

Method

Test Data Description

This study was focused on gender-based DIF and carried out on a population of Caucasians (107,502 females and 75,854 males) who wrote the ACT Assessment (Form 39B) in 1989. Data used were item responses to the Math test. The test included three subtests: Elementary Algebra (24 items, Subtest 1), Algebra/Coordinate Geometry (18 items, Subtest 2), and Plane Geometry/Trigonometry (18 items, Subtest 3). For the total test and for each subtest the mean and reliability (KR-20) for males are somewhat higher than for females (see Table 1).

Insert Table 1 About Here

Based on a sample of 30,000 males and 30,000 females, correlations between subtests 1 and 2, 2 and 3, and 1 and 3 were: .71, .65 and .68 for females and .75, .71, and .74 for males.

Procedure

In this study, four sample sizes (1000/1000, 500/500, 250/250, and 100/100) and two criterion variables (total test score and subtest score) were used for each procedure. To minimize chance results, thirty replications were carried out under each combination of sample size and criterion.

The data analysis procedure for both total test and subtest included the following steps:

1. At each replication the reference group and the focal group with equal number of examinees were randomly selected from the population. Males were identified as the reference group and females as the focal group.

2. The MH chi-square, MH-Z (1/4 MH-Delta), and LR chi-square (uniform and nonuniform) indices were computed for each item using computer program Logreg (Spray 1991) and Mantel (Ackerman 1986). A significance level of .01 was used to identify DIF items.

3. The frequency of identification of DIF was computed over the thirty replications for each item. Three types of DIF were classified: Definite DIF, Probable DIF, and Possible DIF. The absolute mean of MH-Z over the thirty replications was used as the criterion to classify uniform DIF. Based on the sample size of 1000, items for which the absolute mean MH-Z was equal to or greater than .25 were classified as Definite DIF items; items with absolute mean MH-Z between .20 and .25 were classified as Probable DIF items; and items with absolute mean MH-Z greater than .15 but less than .20 were classified as Possible DIF items. The items which met the above criteria must also have been classified as significant at least 15 times at the .01 level in order to be identified as uniform DIF. A nonuniform DIF item would also be inferred if an

item was identified significant at .01 level at least 15 times as having nonuniform DIF.

Results

Uniform DIF. Twelve uniform DIF items were identified using total test score as the criterion. Among these uniform DIF items were four Definite DIF items, four Probable DIF items, and four Possible DIF items. A strong effect of sample size was found on the performance of the MH and LR procedures. For the 12 DIF items identified, the identification rates for each procedure increased sharply as the sample size increased (see Table 2).

Insert Table 2 about here

For Definite DIF items, sample sizes of 500 or larger yielded satisfactory detection rates for items 14, 17, and 58 with both total test and subtest. For example, over 30 replications using total test as the criterion, for a sample size of 250, the detection rate was 25%, on average, for both procedures. When sample sizes of 500 and 1000 were used, the mean percent detection rates increased to 62% and 98% respectively. With subtest score the mean percent detection rates for a sample of 500 were more than 50% for both procedures. Item 19 has the largest amount of DIF with a mean MH-Z of approximately .35, on average, across the four sample sizes. For this item both procedures were able to detect more than 50% for a sample size of 250 with either total test or subtest. The means of MH-Z indices for all items were relatively stable across sample sizes with smaller standard deviations observed for larger sample sizes.

For Probable DIF and Possible DIF items, the sample size of 500 showed extremely low detection rates. For example, with total test score, the average percent detection rates were only

26% for Probable DIF items and 28% for Possible DIF items for the MH procedure and 31% for Probable DIF items and 29% for Possible DIF items for the LR procedure. However, when a sample size of 1000 was used, the mean percent detection rates were 45% higher for probable DIF items and approximately 42% higher for possible DIF items on average for each procedure. With subtest score even lower detection rates were found for all items except item 38, for which the detection rates were greater than 50%. Thus, for Probable DIF and Possible DIF items, to obtain satisfactory results, a sample size of greater than 500 is necessary.

The MH and LR procedures showed very similar detection rates at each sample size and criterion variable with the LR procedure being slightly more powerful. This finding is somewhat different from Swaminathan and Rogers (1990) and Rogers and Swaminathan (1993), who found that the MH procedure had a slight advantage over the LR procedure in identifying items with uniform DIF.

When the effect of the criterion variable was examined, the performance of the two procedures was found to be affected. Using subtest score as the criterion substantially decreased the detection rates and the MH-Z values for Probable DIF and Possible DIF items but the influence was relatively small for Definite DIF items. However, there are three exceptions: items 19, 38, and 58. For item 19 choice of subtest did not affect the detection rates but slightly decreased the MH-Z values. For items 38 and 58 the detection rates slightly increased when subtest was used. Based on our definitions of DIF, items 14 and 17 became Probable DIF; items 4 and 29 became Possible DIF; and items 7, 23, and 32 were no longer DIF. No additional DIF items were identified with subtests.

Of all the 12 DIF items identified using the total test as the criterion, eight DIF items were from subtest 1, no DIF items were found from subtest 2, and four items were from subtest 3. Using total test score as the criterion, the test seemed to favour males slightly with seven items having negative signs and five items having positive signs. However, when subtest was used, the tendency was the opposite. Four items favoured males and five favoured females.

Based on a random sample of 1000 an exploratory factor analysis was carried out on the test using computer program Noharm II (Fraser, 1988). The result suggested that the test had two dimensions, which, however, did not seem related to the subtests. The last ten items (items 51 to 60), including items from three subtests, were highly loaded on factor 2 but not on factor 1, indicating that a speed factor may exist. The finding was supported by the fact that these ten items all had very low difficulty values, p=.22 on average.

Nonuniform DIF. No nonuniform DIF items were found at .01 significant level. However, at .05 level, there were two items (items 33 and 56) for which the LR chi-square statistic was significant 15 times.

False Positives. Type I error rate for each procedure was computed for each sample size and criterion variable. In this study, items for which the mean value of MH-Z was equal to or less than an absolute value of .05 were considered to be free from DIF. Thus, Type I error rate was the proportion of times these items were identified over the 30 replications. For both procedures, the false positives were close to expected values; the LR procedure showed slightly higher rates than the MH procedure.

Discussion

While it is not surprising that increasing sample size increased the power of the MH and LR statistics, the effect on the power of the two procedures produced by sample size for items with different amounts of DIF was noteworthy. For both procedures, to obtain good power in detecting Definite DIF items sample sizes of 500 were necessary and in detecting Probable DIF and Possible DIF items sample sizes of greater than 500 were needed. This finding suggests that one has to take into account the magnitude of DIF in choosing the appropriate sample size for either procedure.

The two procedures demonstrated a high agreement in detecting uniform DIF. An interesting phenomenon is that the LR procedure was slightly more powerful than the MH procedure. Tian et. al. (1994) also showed a similar finding. The results of other studies, however, differ (Swaminathan & Rogers, 1990 and Rogers & Swaminathan 1993). This difference is likely due to the different estimation procedures used in this study.

In terms of choice of criterion, the study yielded findings different from those of previous studies. Clauser et. al. (1991) showed that choice of total test score and subtest score had a substantial effect on DIF detection rates, while Tian et. al. (1994) indicated that choice of total test score and subtest score did not make any difference. In this study, DIF items were identified less frequently with subtest than with total test as the criterion. This effect seemed to be also influenced by size of DIF effect. As shown in the results, the effect was relatively small for those Definite DIF items. For those less obvious DIF items, such as Probable and Possible DIF items, using subtest substantially decreased the power of both procedures. Since the subtests are very short and each subtest measures a distinct subskill, it is likely that these two factors also contributed to the occurrence of this phenomenon.

There is no strong evidence about the existence of nonuniform DIF in the data used. The frequencies so low that they may be due to chance. It would be interesting to know how often nonuniform DIF occurs in real testing situations.

This study provides information on how the MH and LR procedures function with real test data. The findings concerning sample size and criterion effects on DIF identification are likely to be confounded by test length and test dimensionality. Thus, a further question raised in this study relates to the extent that test dimension and test length each affect DIF detection for the MH and LR procedures. It would be interesting to conduct simulated studies in which test dimension and test length can be specified so that the effects of each of the factors on the performance of the MH and LR procedures could be more effectively assessed.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact and item validity from a multidimensional perspective. Journal of Educational Measurement, 29(1), 67-91.

Ackerman, T. A. (1986). Program Man(1: A revised version to incorporate variance estimate of MH-delta statistic and a standardization technique.

Clauser B. E., Mazor K., & Hambleton, R.K. (1991). Influence of the criterion variable on the identification of differentially functioning test items using the Mantel-Haenszel statistic. Applied Psychological Measurement, 15(4),353-359.

Fraser, C. (1988). NOHARM II. A FORTRAN program for fitting unidimensional and multidimensional normal ogive models of latent trait theory. Armidale, Australia: The University of New England, Center for Behavioral Studies.

Hambleton, R. K., Bollwark, J., & Rogers, H. J. (1990). Factors affecting the stability of the Mantel-Haenszel item bias statistic (Rep. No. 203). Amherst: University of Massachusetts, School of Education, Laboratory of Psychometric and Evaluative Research.

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. (Report No. 203). Princeton: Research Statistics Group, Educational Testing Service.

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

Spray, J. (1991). Logreg: A fortran program for calculating uniform and nonuniform DIF indices for the Logistic Regression procedure.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using Logistic Regression procedure. Journal of Educational Measurement, 27(4), 361-370.

Tian, F., Pang, X. L., & Boss, M. W. (1994). The consistency of the Mantel-Haenszel and Logistic Regression DIF identification procedures across sample sizes and over replications. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Table 1

ACT Math Assessment: Descriptive Statistics

	Males			Females		
Criterion	M	SD	KR-20	M	SD	KR-20
Total	28.21	11.69	.92	24.70	10.45	.90
1*	12.82	5.11	.83	11.54	4.78	.81
2	7.41	4.04	.81	6.47	3.70	.76
3	7.99	3.70	.76	6.69	3.25	.68

* The number indicates subtests.

Table 2

Frequencies and MH-Z Values of DIF Items Identified by the MH and LR Procedures Over Thirty Replications Using Total Test and Subtest Scores as Criteria (Alpha Level=.01)

			Frequency				MH-Z			
			Total		Subtest		Total		Subtest	
	Item	Sample	MH	LR	MH	LR	M	SD	M	SD
DEFINITE DIF										
	14 (1)*	1000	**30**	**30**	**24**	**25**	.28	.06	.24	.06
		500	**19**	**20**	**17**	**17**	.31	.10	.27	.09
		250	7	9	4	4	.29	.12	.25	.13
		100	2	**4**	1	3	.31	.24	.25	.23
	17 (3)	1000	**30**	**30**	**29**	**29**	-.29	.06	-.24	.06
		500	**21**	**21**	14	**17**	-.26	.09	-.22	.09
		250	10	10	8	9	-.24	.13	-.20	.14
		100	1	2	2	2	-.22	.23	-.21	.23
	19 (1)	1000	**30**	**30**	**30**	**30**	.35	.07	.34	.06
		500	**24**	**26**	**23**	**23**	.33	.10	.32	.08
		250	**15**	**17**	**16**	**16**	.35	.15	.34	.10
		100	5	6	4	5	.35	.29	.36	.22
	58 (1)	1000	**28**	**29**	**28**	**30**	-.31	.06	-.29	.06
		500	**16**	**16**	**17**	**23**	-.27	.09	-.29	.09
		250	6	9	7	9	-.28	.15	-.32	.15
		100	0	3	2	3	-.33	.27	-.32	.28

Table 2 Continued

		Frequency				MH-Z			
		Total		Subtest		Total		Subtest	
Item	Sample	MH	LR	MH	LR	M	SD	M	SD
PROBABLE DIF									
4 (1)	1000	**19**	**19**	10	12	.21	.07	.17	.06
	500	6	7	2	5	.21	.10	.16	.09
	250	3	4	2	2	.20	.15	.17	.15
	100	0	1	0	1	.20	.24	.18	.25
29 (1)	1000	**20**	**24**	10	12	.21	.04	.16	.04
	500	4	6	1	1	.16	.08	.13	.08
	250	0	1	0	0	.14	.11	.11	.10
	100	1	0	0	0	.17	.24	.11	.23
38 (1)	1000	**23**	**24**	**24**	**27**	-.21	.07	-.24	.06
	500	9	12	**16**	**17**	-.20	.09	-.23	.09
	250	4	1	3	6	-.21	.12	-.23	.11
	100	2	2	1	1	-.22	.25	-.24	.23
52 (1)	1000	**23**	**25**	**23**	**25**	-.20	.06	-.20	.07
	500	12	12	12	13	-.20	.08	-.20	.09
	250	4	4	4	4	-.17	.13	-.17	.13
	100	1	1	0	1	-.19	.18	-.19	.15

Table 2 Continued

		Frequency				MH-Z			
		Total		Subtest		Total		Subtest	
Item	Sample	MH	LR	MH	LR	M	SD	M	SD
POSSIBLE DIF									
7 (3)	1000	**17**	**18**	10	11	-.19	.06	-.14	.07
	500	7	8	4	5	-.18	.08	-.14	.09
	250	2	4	1	1	-.17	.13	-.11	.14
	100	0	1	0	0	-.19	.21	-.13	.22
9 (1)	1000	**18**	**20**	**15**	**15**	.19	.07	.17	.07
	500	7	7	5	5	.17	.11	.16	.10
	250	8	5	4	4	.18	.17	.17	.16
	100	1	2	1	2	.15	.21	.14	.26
23 (3)	1000	**22**	**23**	8	7	-.18	.07	-.13	.07
	500	11	12	3	4	-.19	.08	-.14	.09
	250	4	6	2	4	-.21	.12	-.17	.14
	100	0	1	0	1	-.19	.18	-.14	.20
32 (3)	1000	**24**	**25**	10	10	-.19	.05	-.14	.06
	500	8	8	4	4	-.17	.08	-.12	.08
	250	3	1	1	1	-.16	.15	-.10	.14
	100	1	4	1	2	-.16	.19	-.18	.21

* The number in the bracket indicates the number of subtests.