Research Report

DIF Detection With Small Samples: Applying Smoothing Techniques to Frequency Distributions in the Mantel-Haenszel Procedure

Lei Yu

Tim Moses

Gautam Puhan

Neil Dorans

August 2008

ETS RR-08-44

DIF Detection With Small Samples: Applying Smoothing Techniques to Frequency Distributions in the Mantel-Haenszel Procedure

Lei Yu, Tim Moses, Gautam Puhan, and Neil Dorans
ETS, Princeton, NJ

August 2008

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS' constituents and the field.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

All differential item functioning (DIF) methods require at least a moderate sample size for effective DIF detection. Samples that are less than 200 pose a challenge for DIF analysis. Smoothing can improve upon the estimation of the population distribution by preserving major features of an observed frequency distribution while eliminating the noise brought about by irregular data points. This study applied smoothing techniques to frequency distributions and investigated the impact of smoothed data on the Mantel-Haenszel (MH) DIF detection in small samples. Eight sample-size combinations were randomly drawn from a real data set to make the study realistic and were replicated 80 times to produce stable results. The population DIF results were used as the criteria to evaluate sample estimates using root-mean square difference (RMSD), bias analysis, and Type II error rate. Loglinear smoothing was found to provide slight to moderate improvements in MH DIF estimation with small samples.

Key words: DIF, small samples, Mantel-Haenszel, loglinear smoothing, bias

Table of Contents

List of Tables

Page

List of Figures

Page

Introduction

Differential item functioning (DIF) examines conditional item performance across groups and DIF analysis has become a standard operational procedure for many testing programs. All DIF methods require at least a moderate sample size for effective DIF detection. For example, the Mantel-Haenszel (MH) method (Holland & Thayer, 1988), one of the most popular DIF detection procedures with a known advantage of working effectively with small number of examinees, requires a sample size of 200 to be adequate (Mazor, Clauser, & Hambleton, 1992). In real testing situations, however, small samples (less than 200) occur, which poses a challenge for DIF analysis.

Research on detecting DIF in small samples is limited. Parshall and Miller (1995) proposed an exact test as an approach preferred over the standard asymptotic procedure for DIF analysis using the MH method on small samples and found that the two methods produced very similar results. Roussos and Stout (1996) studied the Type I error inflation with small samples for the MH and the simultaneous item bias test (SIBTEST) methods and showed that the Type I error rate did not differ much between the two procedures. Zwick, Thayer, and Lewis (1997, 1999, 2000) and Zwick and Thayer (2002) developed an empirical Bayes (EB) enhancement of the MH DIF method and found that the EB methods using the same priors produced improved DIF estimates over the standard MH approach, especially for small samples. Sinharay et al. (2006) applied a full Bayesian approach using different prior information for different item types in the MH DIF analysis with small samples and found that the Bayesian approach improved DIF estimation over other existing methods. The Bayesian approach may be a promising option in dealing with the small sample issue, but it does require the accumulation of enough past data that can be used to specify the prior information.

Sample size is important in attaining statistical precision. When sample sizes are small, irregularities in score frequency distributions are likely to occur. The number of test-takers with a given score may not change gradually as the scores increase. Instead, the numbers may fluctuate. Such irregularities cause problems, as they may not be generalizable to other groups of examinees. Therefore, DIF results obtained with one small sample may not be replicable when another small sample is used.

A population distribution can be estimated based on an observed sample distribution. Smoothing an observed distribution improves upon the process of estimating the population

distribution, as smoothing is intended to preserve the major features of an observed frequency distribution while eliminating the noise brought about by irregular data points due to sampling. Loglinear smoothing and kernel smoothing are two major smoothing techniques. Loglinear smoothing smoothes the data based on loglinear models by preserving a number of moments (e.g., mean, variance, skewness, or kurtosis) in the smoothed distribution, using the same values as those obtained in the observed distribution. Kernel smoothing refers to a class of functions that use different weighting schemes to compute local averages. The Gaussian kernel function, one of the functions that can be employed with kernel smoothing, uses the Gaussian distribution (commonly known as the normal distribution) as the weighting function. With this function, scores around a particular evaluation point receive most of the weight (Lyu, Dorans, & Ramsey, 1995). Both loglinear smoothing and kernel smoothing were used in the current study to smooth the frequency distributions of examinee responses by group.

The Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988) is a well-established, practical, and powerful method for detecting DIF. It estimates the ratio of the probabilities of the reference and the focal groups answering an item correctly using K x 2 x 2 contingency tables. The current study applied smoothing techniques to frequency distributions and investigated the impact of using smoothed data on MH DIF detection in small samples. DIF analysis was conducted on three types of data: unsmoothed or raw data, loglinear-smoothed data and kernel-smoothed data. Varying sample sizes were randomly drawn from a real data set to make the study as realistic as possible. Each sample size condition was replicated 80 times to produce stable results. DIF results based on the complete data set were used as the criterion to evaluate the results obtained from the samples using root-mean square difference (RMSD), bias analysis, and Type II error rate.

Method

Sample

A large admission test that was formula scored was used in the study. There were five options for each item. The population consisted of 47,686 examinees, of whom 11,910 were Asians and 21,494 were Whites. In the Asian-White comparison, eight items were detected as demonstrating significant DIF, and all of them were used in the current study. An additional item, one that contained the least amount of DIF among all the items of the test, was also included. This minimum DIF item provided an examination of DIF detection in small samples

from a perspective that was different from what was provided by the other DIF items. Realistic samples of the real data were randomly drawn with replacement from the population for eight sample size combinations of the focal and reference groups, which included 50/300, 75/300, 100/300, 150/300, 200/300, 300/400, 300/700, and 700/2100. Each sample size combination was replicated 80 times for stability of results.

Mantel-Haenszel Procedure

The Mantel-Haenszel procedure, applied to DIF analysis by Holland and Thayer (1988), studies performance differences between matched groups on dichotomously scored items. The procedure compares the ratio of the probabilities of two groups answering an item correctly across all score levels (K). The obtained estimate is known as the odds ratio. In this procedure, data are constructed as a K x 2 x 2 contingency table. Table 1 presents an example of a 2 x 2 contingency table at a given score level k. Such a table is required for each score level, resulting in a K x 2 x 2 contingency table to be used in the MH procedure.

Table 1

An Example of a 2 x 2 Contingency Table at Score Level K

		Item score		
		1	0	Total
Group	Focal	R_{Fk}	W_{Fk}	N_{Fk}
	Reference	R_{Rk}	W_{Rk}	N_{Rk}
	Total	R_{Tk}	W_{Tk}	N_{Tk}

The common odds ratio is computed using the following formula:

$$a_{MH} = \frac{\sum_k R_{Rk} W_{Fk} / N_{Tk}}{\sum_k R_{Fk} W_{Rk} / N_{Tk}}.$$

R_{Rk} and R_{Fk} represent respectively the number of examinees in the reference and the focal groups answering an item correctly at score level k; W_{Rk} and W_{Fk} denote respectively the number of the examinees in the reference and focal groups answering an item incorrectly at that score level; N_{Tk} refers to the total number of the examinees at the same score level. As mentioned earlier, the data used in the study came from a formula-scored test with 5-choice items. Formula score is a scoring

method that corrects the total score for wrong answers. In this study, correct responses were scored as 1, incorrect response as -.25, and omitted or not-reached responses as 0.

The null hypothesis test associated with the MH procedure is a chi-square test, which tests the hypothesis that the focal and the reference groups at a given k level of ability have the same odds of answering the item correctly across all K levels. It is expressed symbolically as follows

$$H_0: \ \frac{R_{Rk}}{W_{Rk}} = \frac{R_{Fk}}{W_{Fk}} \quad k = 1, \ldots, K.$$

The alternative hypothesis is

$$H_a: \ \frac{R_{Rk}}{W_{Rk}} \neq \frac{R_{Fk}}{W_{Fk}} \quad k = 1, \ldots, K,$$

or

$$H_a: \ \frac{R_{Rk}}{W_{Rk}} = a\frac{R_{Fk}}{W_{Fk}} \quad k = 1, \ldots, K \text{ and } a \neq 1,$$

and when $a = 1$, the alternative hypothesis is equal to the null hypothesis.

The odds ratio provides an estimate of DIF effect size. To facilitate interpretation, the MH DIF statistic is transformed onto the ETS delta scale of difficulty by taking its natural log and multiplying the result by –2.35:

$$MH\ D-DIF = -2.35 \ln(a_{MH}).$$

Negative values indicate that the item is more difficult for the focal group, and positive values indicate that the item is more difficult for the reference group.

As part of the MH procedure, significance testing is also carried out. The absolute value of MH D-DIF and the significance test are used jointly to determine the level of severity of DIF, classifying all items into C-, B-, A, B+, and C+ categories using the ETS classification rules. An item is identified as showing significant amount of DIF and is assigned the category of C if MH

D-DIF is significantly different from 1 in absolute value and the absolute value is greater than or equal to 1.50.

Smoothing Procedures

Two smoothing procedures—loglinear smoothing and kernel smoothing—were used in this study. Loglinear smoothing fits a loglinear model to the observed discrete data using maximum likelihood estimation (Holland & Thayer, 1987; 2000). The following formula is used for loglinear smoothing:

$$\log_e(p_k) = \alpha + \beta_1 x_k + \beta_2 x_k^2 + \beta_3 x_k^3 \ ... \ + \beta_i x_k^i,$$

where x is the test score with possible values of $0, k, ... K$. p_k is the probability of obtaining test score x_k. α is a constant that restricts the sum of all probabilities to 1. $\beta_1, \beta_2, \beta_3$ and β_i are parameters to be estimated in order to preserve in the smoothed distribution the moments obtained from the observed distribution. For example, if $I = 3$ $(i = 1, 2, ...I)$, the loglinear model will preserve three moments: mean, standard deviation, and skewness of the observed distributions in the smoothed distribution. The estimated probabilities $(\hat{p}_k)$ need to meet this condition:

$$\sum_k x_k^i \hat{p}_k = \sum_k x_k^i (\frac{n_K}{N}),$$

where n_k is the number of examinees at score level k and N is the total number of examinees. Specifically, the probability for the mean, standard deviation, and skewness will be estimated using the following:

$$\sum_k x_k^1 \hat{p}_k = \sum_k x_k^1 (\frac{n_k}{N})$$
$$\sum_k x_k^2 \hat{p}_k = \sum_k x_k^2 (\frac{n_k}{N})$$
$$\sum_k x_k^3 \hat{p}_k = \sum_k x_k^3 (\frac{n_k}{N}).$$

The test used in this study was formula scored, and the scores were rounded to the nearest integer. This rounding of formula scores produces low frequencies at regular intervals known as *teeth*, and they would occur at exactly the same score levels in a different sample of examinees. An example of a score distribution of a formula-scored test is presented in Figure 1. The teeth feature of the data complicates the smoothing process, however. Loglinear smoothing makes it possible to smooth the teeth separately from other score levels by assigning them to a *subset selection vector*. Therefore, loglinear smoothing smoothes the teeth in the distribution as well as the main distribution. Three overall moments and two moments of the teeth distributions were preserved throughout. The likelihood ratio chi-square statistics of overall fit were generally close to the degrees of freedom for the largest sample sizes considered in this study. This suggests that this fixed model was reasonable for most considered conditions, though somewhat overparameterized for the smallest sample sizes considered.

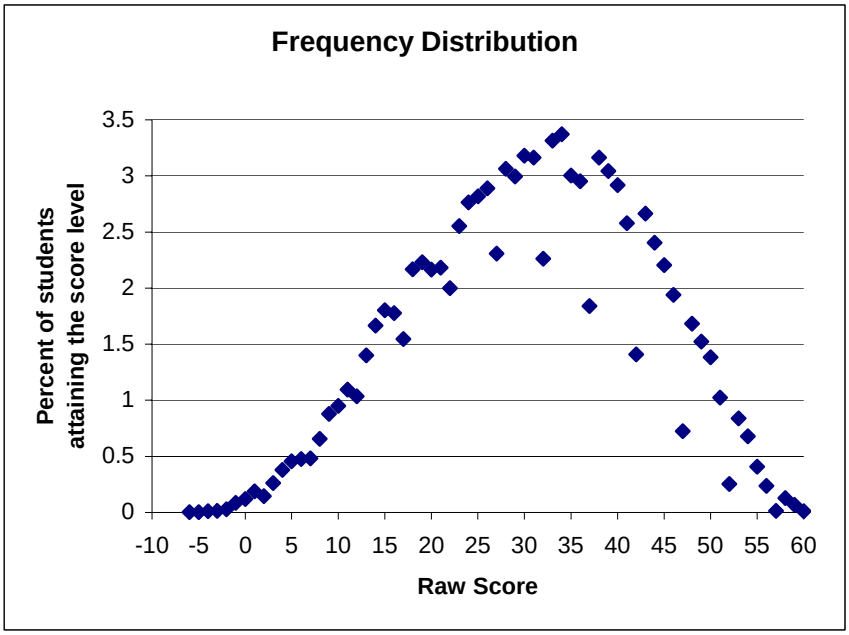


Figure 1. **Score distribution of a formula-scored test.**

Kernel smoothing uses weighting functions to compute local averages. The four frequency distributions (i.e., frequency distributions of rights and wrongs for the reference and focal groups) were kernel smoothed separately. For instance, kernel smoothing of the frequency

distribution of the reference examinees answering the item correctly (R_{Rk} for all K levels) is as follows:

$$KSR_{Rk} = \sum_j w_j(x_k) R_{Rj},$$

where

$$w_j(x_k) = \frac{\exp\left[-\left(\frac{x_j - x_k}{\sigma_x}\right)^2 \left(\frac{1}{2h}\right)\right]}{\sum_i \exp\left[-\left(\frac{x_i - x_k}{\sigma_x}\right)^2 \left(\frac{1}{2h}\right)\right]},$$

where $w_j(x_k)$ is the weight applied to score x_k, and h is the kernel bandwidth parameter set based on the reference sample who got the item right, $\boldsymbol{h} = 1.1\left(\sum_{\boldsymbol{K}} \boldsymbol{R}_{\boldsymbol{RK}}\right)^{-.2}$. Similar relationships hold for the other three groups, including the reference sample who got it wrong, the focal sample who got it right, and the focal sample who got it wrong..

Adjustments were made to account for conditions that were not possible (such as the maximum score on the matching variable for the students who did not get the item right). The final kernel-smoothed rights and wrongs of the reference and focal groups would sum to the observed total sample sizes.

Evaluation of Results

Three sets of MH D-DIF statistics were produced for each item for various sample size combinations for three types of data: one for unsmoothed or raw data, one for loglinear-smoothed data, and one for kernel-smoothed data. The sample estimates were compared to the MH D-DIF values obtained from the population consisting of all examinees. The population values were used as the criteria to evaluate the sample estimates using three methods: RMSD, bias, and Type II error rate.

RMSD. Root-mean square difference was computed between the average of sample estimates at a given condition and the criterion using the formula:

$$RMSD(MH\ D-DIF)=\sqrt{\frac{1}{80}\sum_{N=1}^{80}(\hat{\alpha}_{NMH}-\alpha)^2}$$

$$RMSD(MHLS\ D-DIF)=\sqrt{\frac{1}{80}\sum_{N=1}^{80}(\hat{\alpha}_{NLS}-\alpha)^2}$$

$$RMSD(MHKS\ D-DIF)=\sqrt{\frac{1}{80}\sum_{N=1}^{80}(\hat{\alpha}_{NKS}-\alpha)^2}\ ,$$

where $\hat{\alpha}_N$ represents the sample estimate for item N using a given method and α represents the criterion value. The larger the RMSD, the more the sample estimate deviates from the criterion value. The smaller the RMSD, the closer the sample estimate is to the criterion value.

Bias. Sample estimates, $\hat{\alpha}_N$, are used to estimate the criterion value, and how far the average value of the sample estimates differ from the criterion value, known as bias, can also be evaluated. Bias is the expected difference between the average estimate and the criterion. It was estimated in the current study using the following:

$$Bias_{MH}=\left(\frac{1}{80}\sum_{N=1}^{80}\hat{\alpha}_{NMH}\right)-\alpha$$

$$Bias_{MHLS}=\left(\frac{1}{80}\sum_{N=1}^{80}\hat{\alpha}_{NLS}\right)-\alpha$$

$$Bias_{MHKS}=\left(\frac{1}{80}\sum_{N=1}^{80}\hat{\alpha}_{NKS}\right)-\alpha\ .$$

While RMSD can take on only positive values, bias can be positive or negative. Positive values indicate that, on average, sample values overestimate the population or criterion value. Negative values indicate that, on average, sample values underestimate the population value.

Type II error rate. A Type II error occurs when an item that is flagged as C DIF in the population fails to be identified as having C DIF in a particular combination of focal and reference groups. Eight items included in this study contained known true large DIF. Identifying the rate at which they were not detected in various combinations of sample sizes was important.

Type II error rate was calculated in a particular sample combination by dividing the number of times where DIF was not identified by the total number of replications, which was 80.

Results

Included in the study were eight C DIF items and one item with no DIF identified in the Asian-White comparison in the population. The Asian was the focal group and the White was the reference group. Among the eight DIF items, four were negative DIF, indicating that the items were unexpectedly more difficult for the focal group, conditioning on the total score; four were positive DIF, indicating that the items were unexpectedly easier for the focal group, conditioning on the total score. Using the absolute value of 1.5 as the criterion, the amounts of DIF in these items also varied, with some containing larger amount of DIF (e.g., Item 10) and some smaller (e.g., Item 6). Characteristics of these items from the population analysis are presented in Table 2. Delta values and r-biserial correlations are based on the total of the focal and reference groups.

Table 2

Summary of Studied Items

Category	Item	MH D-DIF	Delta	R-biserial
Negative DIF	6	-1.51	7.64	0.47
	7	-2.63	9.34	0.59
	10	-2.96	9.49	0.55
	12	-1.64	12.40	0.54
No DIF	49	0.06	11.99	0.38
Positive DIF	4	1.66	11.46	0.50
	13	2.81	13.71	0.43
	57	1.63	12.50	0.45
	59	1.81	13.20	0.46

Note. MH = Mantel-Haenszel.

Sample Estimates, Root-Mean Square Difference, and Bias

The MH D-DIF estimates over the 80 replications are summarized in box-and-whisker plots, which provide a method for summarizing data measured on an interval scale using the median, upper and lower quartiles, minimum, and maximum values. Take, for example, the plot at the left bottom of Figure 2 for Item 4 at the focal sample size of 75 and the reference sample

size of 300. The box itself contains the middle 50% of the data. The upper line of the box represents the upper or the third quartile, which 75% of the values fall at or below, and the lower line of the box represents the lower or the first quartile, which 25% of the values fall at or below. The box length that covers these two quartiles is known as the interquartile range (IQR) and indicates sample variability. The line inside the box is the median or the middle of the distribution, with half of the estimates above it and half below. The mean of the distribution is symbolized with a "+" sign. The vertical lines that extend from the two sides of the box are whiskers, which end with cross bars when the minimum and maximum values of the estimates are reached and there are no outliers. When outliers are present, the whiskers are extended to a value of 1.5 times the interquartile range. Outliers are represented separately by the small boxes beyond the whiskers. The position of the box in its whiskers and the position of the line in the box also provide information on whether the sample is symmetric or skewed. In a boxplot, the whiskers have approximately the same length as the box or can be slightly longer for a sample from a normally distributed population.

The results of MH estimates for each item across all the eight sample size combinations ordered from the largest to the smallest are presented in Figures 2-10. Each figure consists of eight plots, and each plot contains three boxplots of estimates: one for the original unsmoothed or raw data (MH), one for loglinear-smoothed data (MHLS), and one for the kernel-smoothed data (MHKS). The Y-axis represents MH D-DIF estimates. The horizontal dotted line represents the criterion value, which is the DIF result from the population and is used to evaluate sample estimates. The graphs show that, in general, with the decreases in sample size combinations, the boxplots become bigger, indicating that the ranges of the middle of the 50% of the sample estimates become wider. This was expected and true for all items for smoothed and unsmoothed estimates. Boxplots of MH and MHLS were similar in shape. In fact, MHLS estimates are less variable and therefore demonstrate an advantage over MH estimates, especially at the smallest sample size combinations of 50/300 and 75/300. The MHKS results show a different pattern. The center of the boxplots tends to deviate from the criterion and from both MH and MHLS results, especially for sample size combinations where focal samples were above 200.

RMSDs between the average sample estimates over 80 replications using different types of data for the various sample size combinations are summarized in Table 3. Sample size combinations are arranged from large to small for each item. Smaller RMSD indicates that

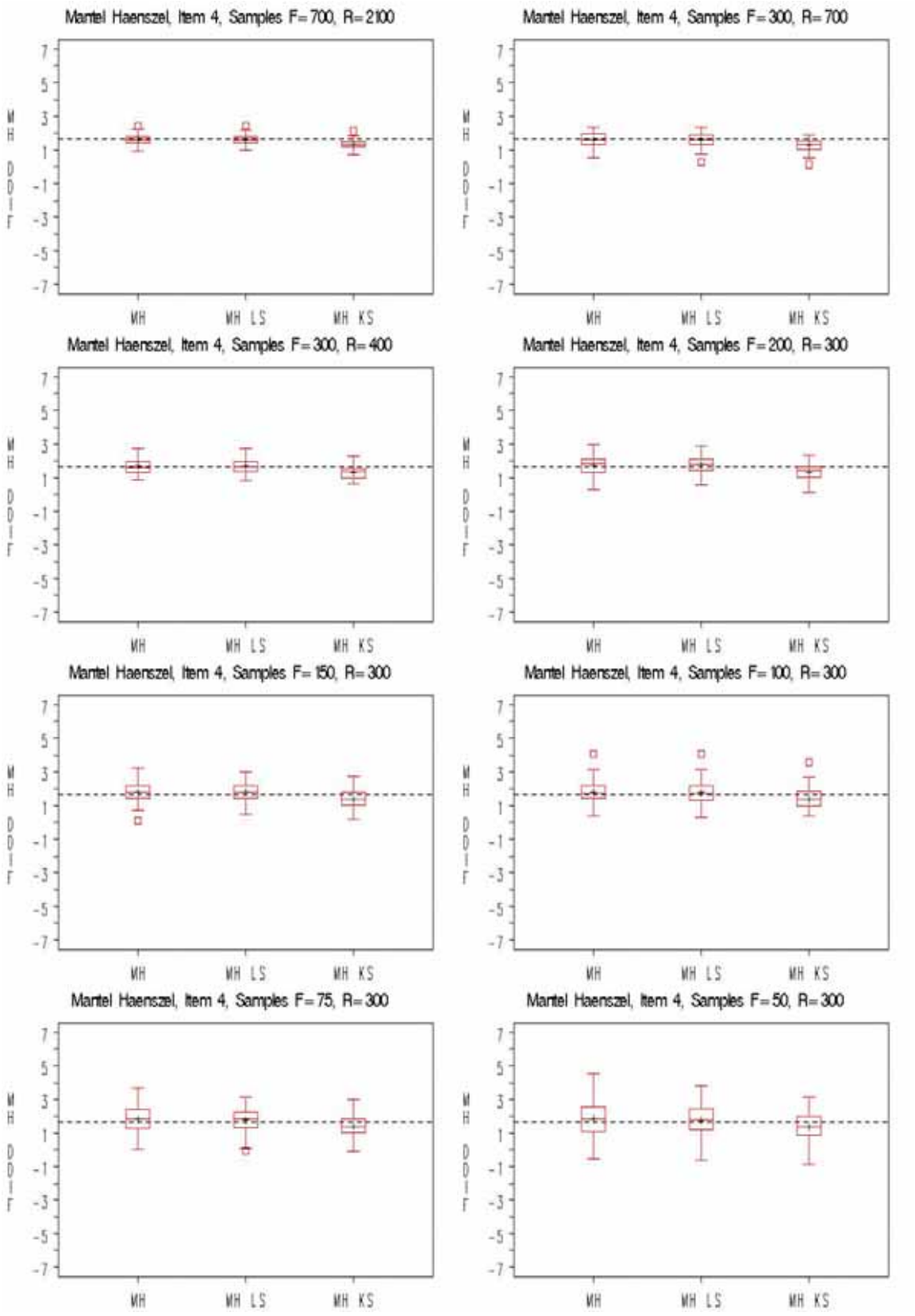


***Figure 2*. Boxplots for Item 4.**

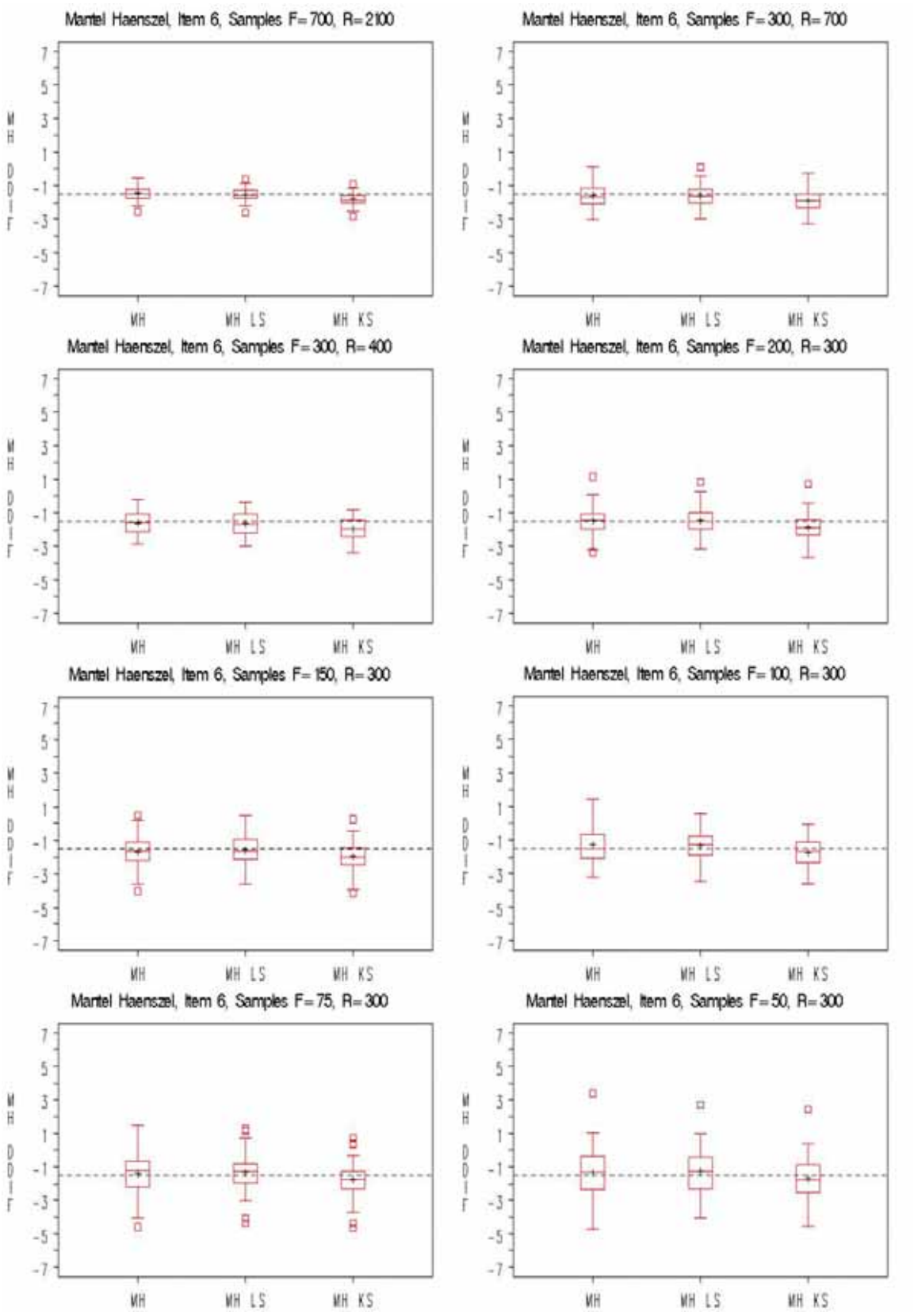


***Figure 3*. Boxplots for Item 6.**

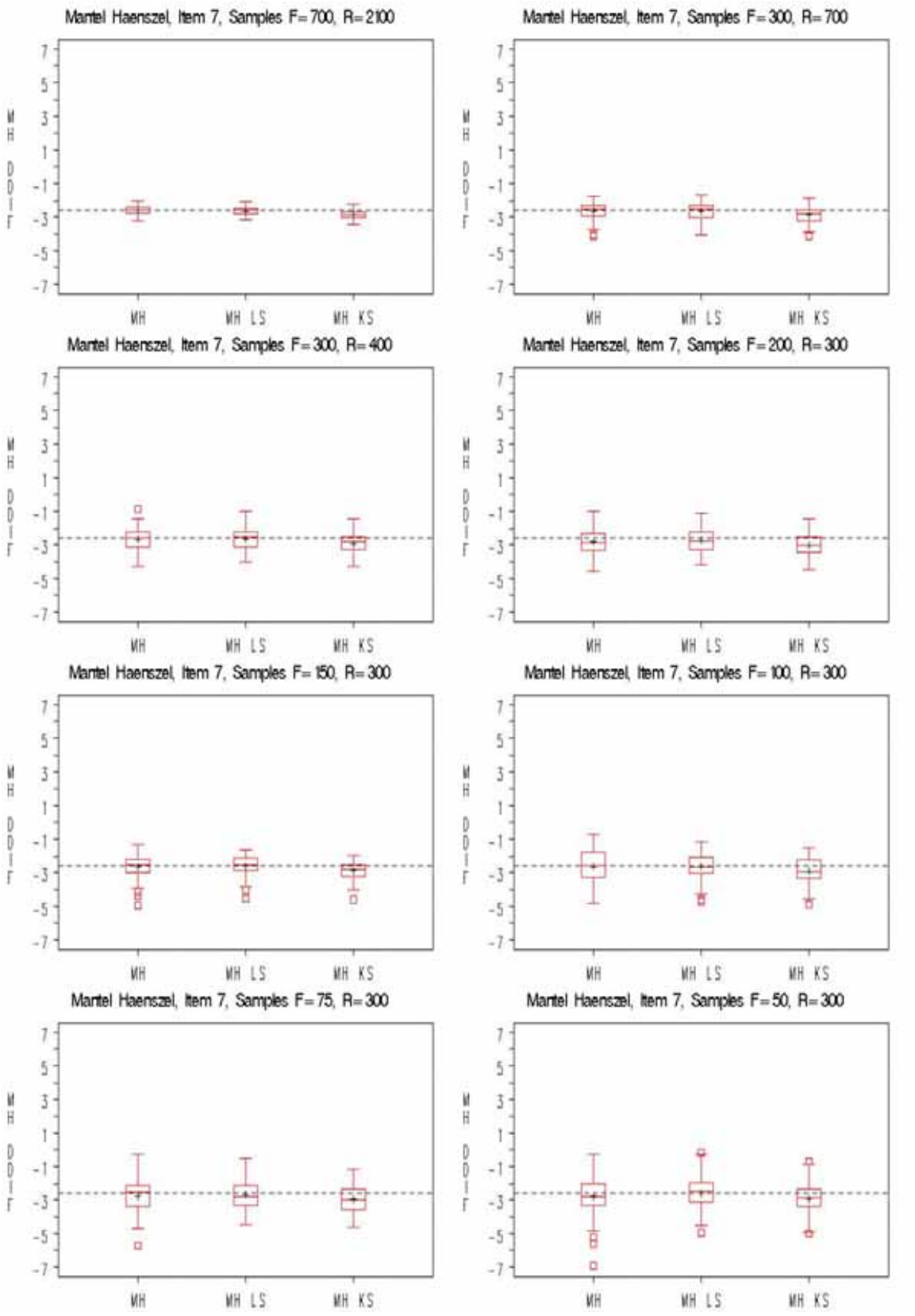


***Figure 4*. Boxplots for Item 7.**

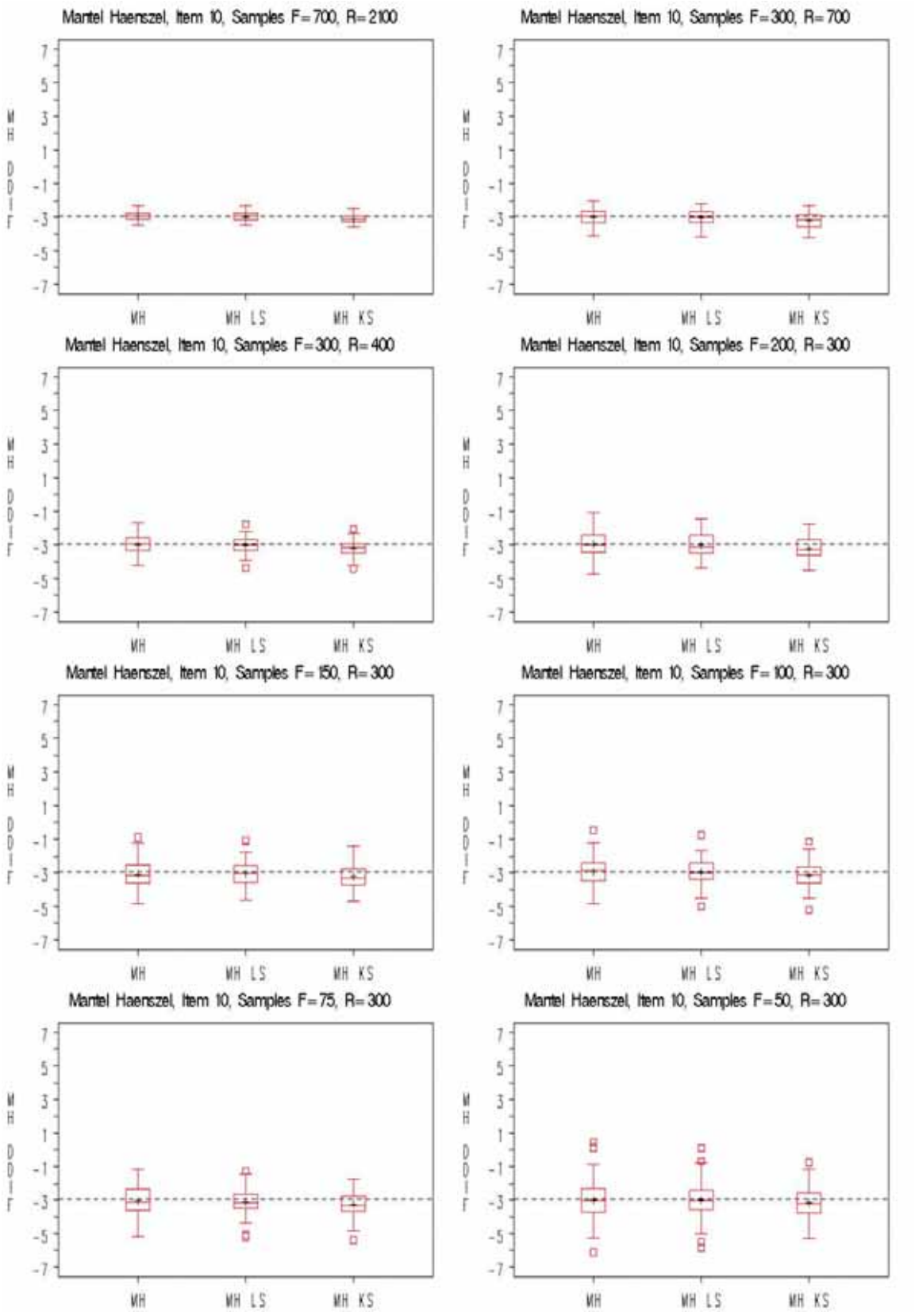


***Figure 5*. Boxplots for Item 10.**

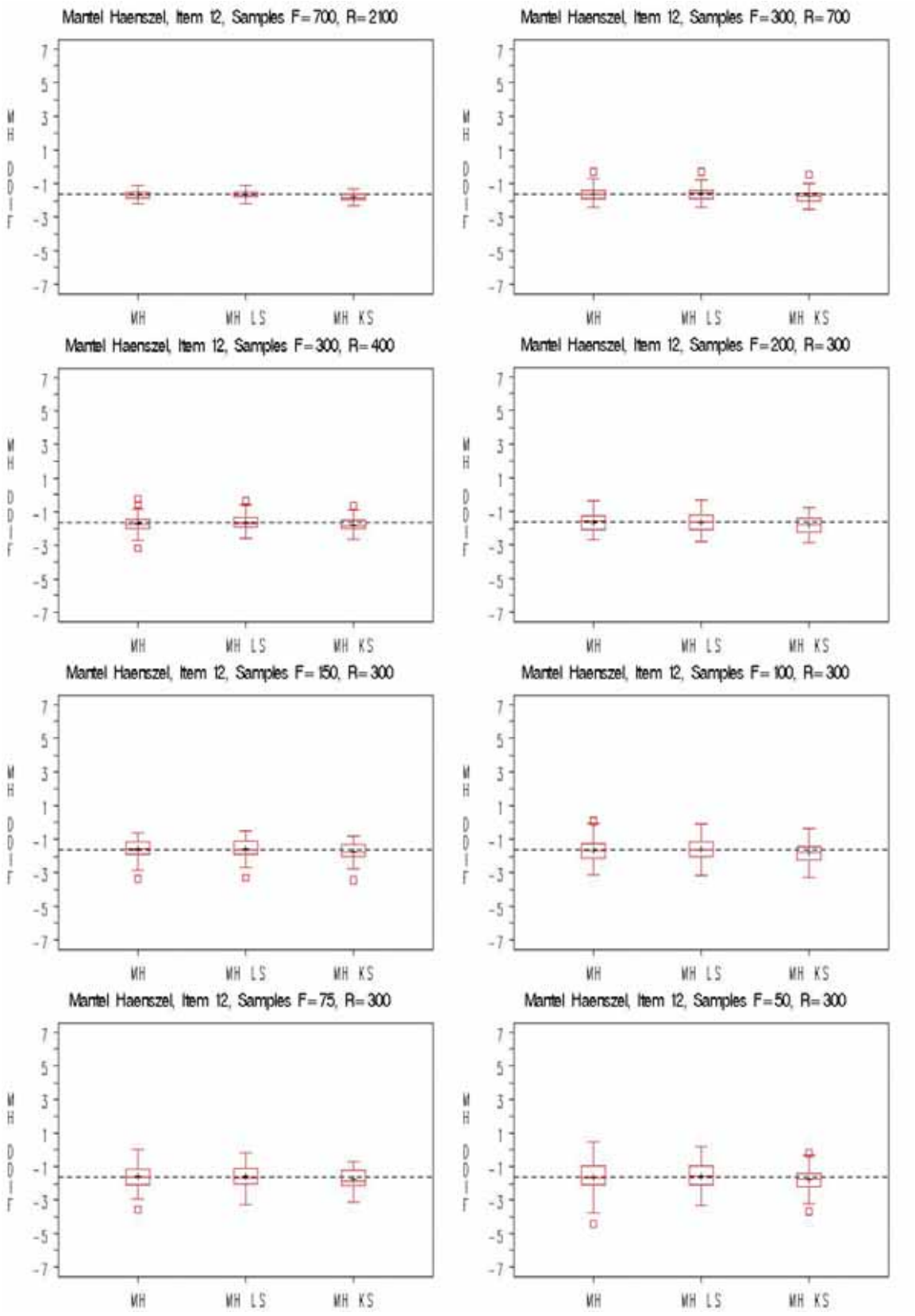


***Figure 6*. Boxplots for Item 12.**

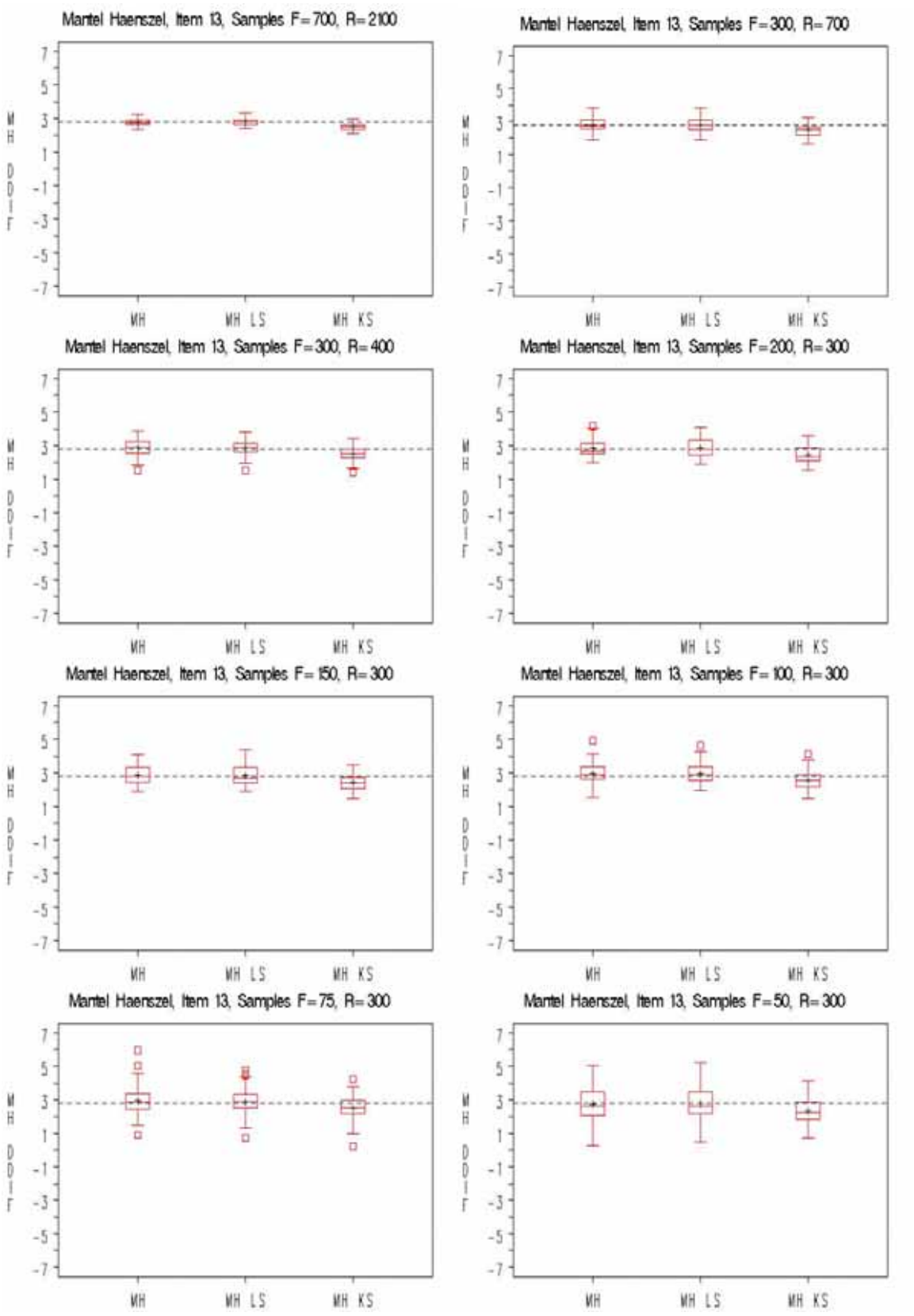


***Figure 7*. Boxplots for Item 13.**

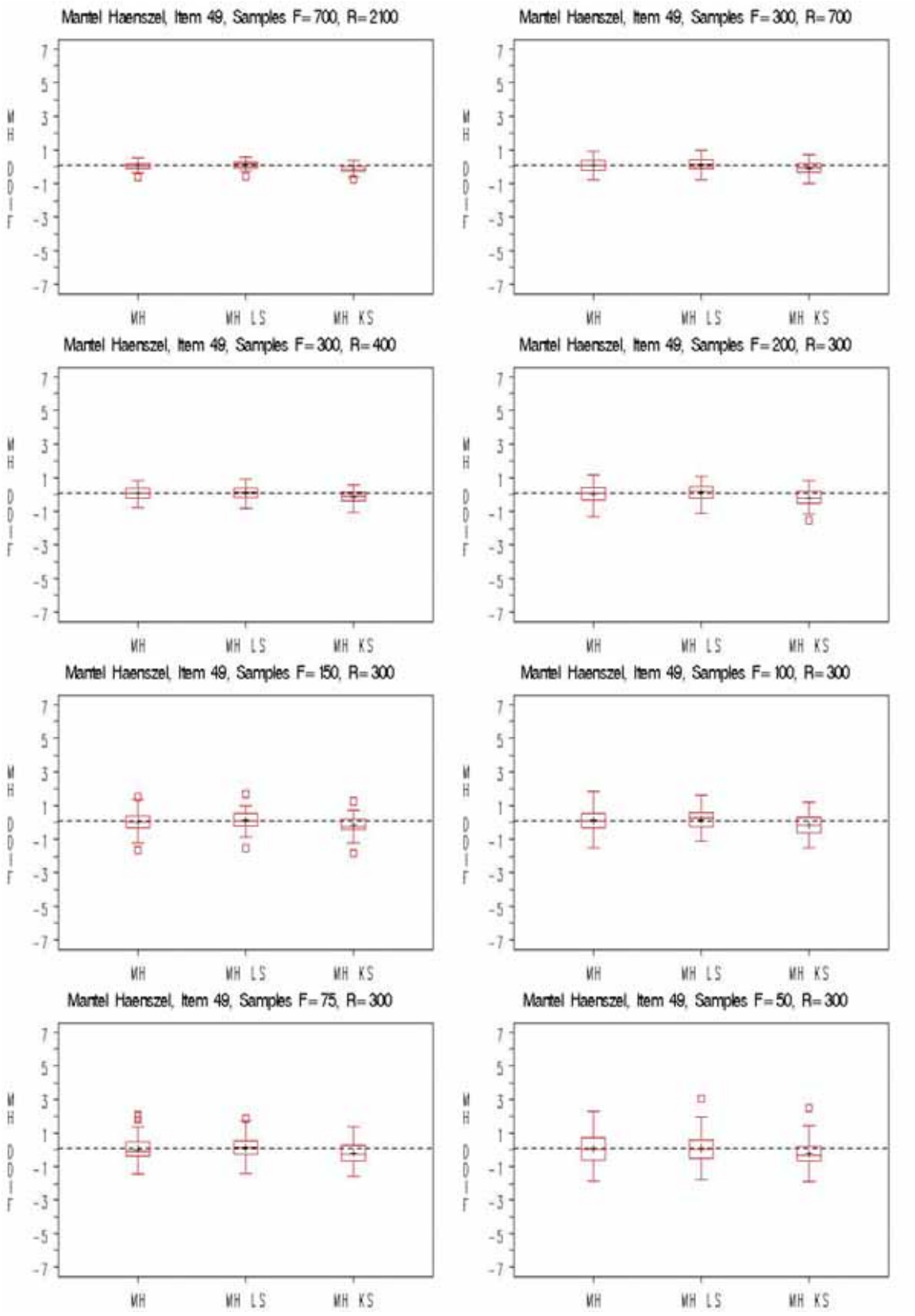


***Figure 8*. Boxplots for Item 49.**

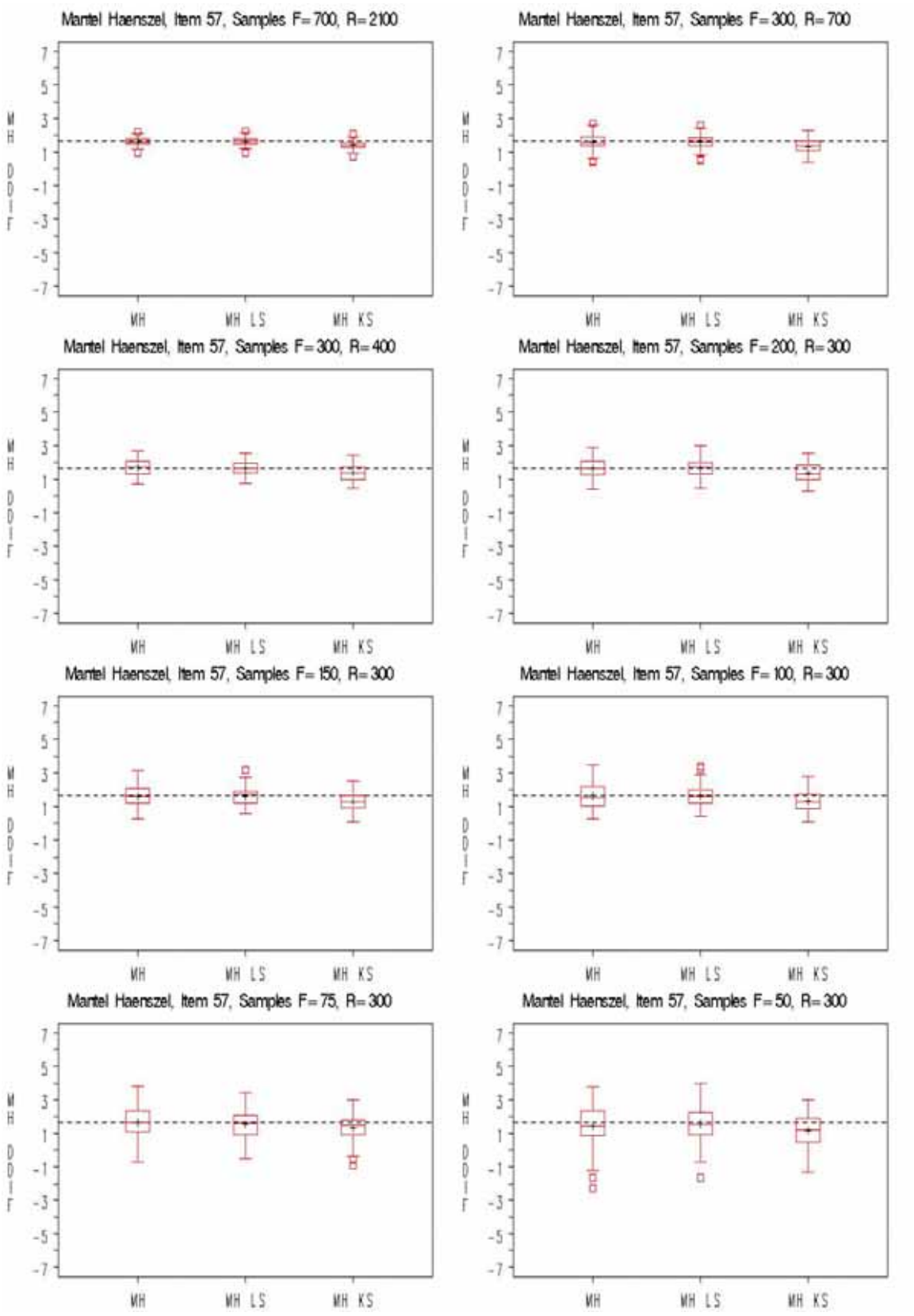


***Figure 9*. Boxplots for Item 75.**

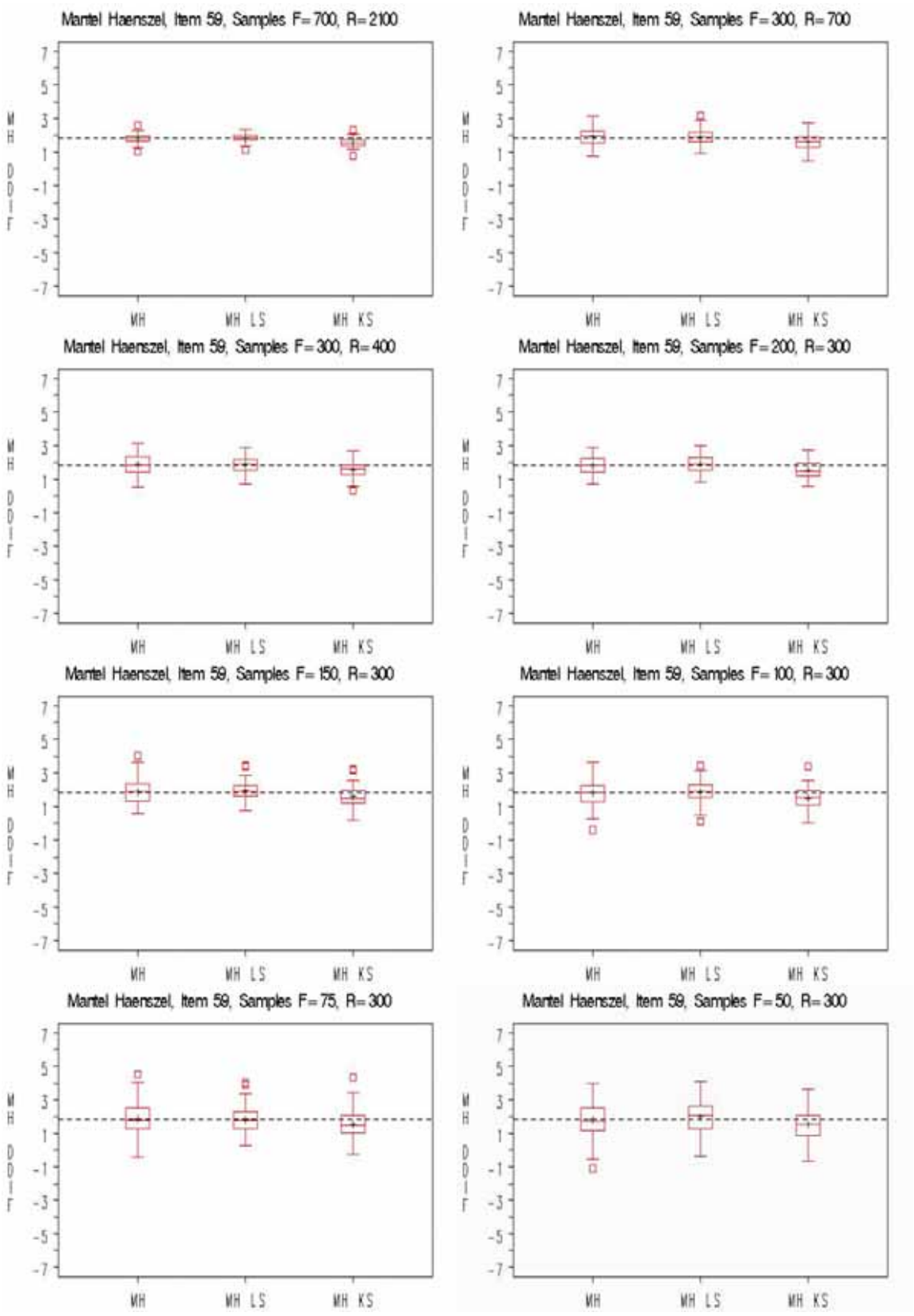


***Figure 10*. Boxplots for Item 59.**

Table 3

Root-Mean Square Difference and Bias Estimates for All Items

		Size		RMSD			Bias		
Item	Criterion	Focal	Reference	MH	MHLS	MHKS	MH	MHLS	MHKS
4	1.66	700	2,100	0.29	0.28	0.42	-0.05	-0.06	-0.32
		300	700	0.41	0.41	0.54	-0.07	-0.07	-0.39
		300	400	0.46	0.43	0.52	0.02	0.03	-0.34
		200	300	0.57	0.54	0.58	0.07	0.09	-0.33
		150	300	0.59	0.58	0.60	0.12	0.13	-0.28
		100	300	0.66	0.67	0.65	0.13	0.09	-0.28
		75	300	0.78	0.71	0.72	0.17	0.10	-0.27
		50	300	1.12	0.92	0.91	0.18	0.08	-0.30
6	-1.51	700	2,100	0.36	0.34	0.45	0.02	-0.03	-0.31
		300	700	0.64	0.59	0.71	-0.08	-0.11	-0.43
		300	400	0.68	0.64	0.77	-0.09	-0.13	-0.46
		200	300	0.79	0.74	0.79	0.04	0.03	-0.35
		150	300	0.96	0.83	0.94	-0.18	-0.06	-0.46
		100	300	1.07	0.92	0.89	0.20	0.16	-0.25
		75	300	1.21	1.08	1.05	0.12	0.17	-0.26
		50	300	1.44	1.30	1.26	0.15	0.24	-0.20
7	-2.63	700	2,100	0.25	0.25	0.35	0.00	-0.03	-0.26
		300	700	0.51	0.51	0.55	-0.02	-0.03	-0.28
		300	400	0.64	0.63	0.68	-0.07	-0.06	-0.32
		200	300	0.75	0.70	0.76	-0.18	-0.13	-0.40
		150	300	0.69	0.59	0.61	-0.02	0.02	-0.28
		100	300	0.94	0.79	0.78	0.00	0.00	-0.29
		75	300	1.03	0.87	0.87	-0.15	-0.07	-0.34
		50	300	1.30	1.02	0.97	-0.15	0.02	-0.32
10	-2.96	700	2,100	0.26	0.25	0.28	0.02	-0.01	-0.14
		300	700	0.47	0.47	0.51	-0.06	-0.10	-0.26
		300	400	0.48	0.46	0.51	-0.03	-0.08	-0.25
		200	300	0.75	0.65	0.66	-0.02	-0.07	-0.26
		150	300	0.80	0.75	0.76	-0.13	-0.09	-0.28
		100	300	0.85	0.76	0.75	0.01	-0.01	-0.21
		75	300	0.89	0.78	0.77	-0.12	-0.14	-0.30
		50	300	1.19	1.02	0.91	-0.05	-0.04	-0.21
12	-1.64	700	2,100	0.24	0.24	0.27	-0.04	-0.03	-0.16
		300	700	0.43	0.42	0.40	0.02	0.04	-0.10
		300	400	0.48	0.44	0.43	-0.05	0.01	-0.13
		200	300	0.52	0.53	0.52	-0.02	-0.02	-0.19
		150	300	0.56	0.55	0.51	0.03	0.06	-0.12
		100	300	0.68	0.67	0.63	-0.02	0.03	-0.15

Table continues

Table 3 (continued)

		Size		RMSD			Bias		
Item	Criterion	Focal	Reference	MH	MHLS	MHKS	MH	MHLS	MHKS
		75	300	0.68	0.67	0.59	0.04	0.07	-0.12
		50	300	0.91	0.80	0.71	0.02	0.07	-0.16
13	2.81	700	2,100	0.22	0.22	0.38	-0.04	-0.04	-0.32
		300	700	0.45	0.44	0.52	0.00	0.00	-0.33
		300	400	0.49	0.44	0.50	0.08	0.07	-0.30
		200	300	0.55	0.54	0.59	0.03	0.05	-0.34
		150	300	0.60	0.60	0.62	0.05	0.03	-0.37
		100	300	0.66	0.57	0.58	0.13	0.14	-0.26
		75	300	0.87	0.78	0.75	0.13	0.10	-0.32
		50	300	1.02	0.93	0.93	-0.06	-0.04	-0.49
49	0.06	700	2,100	0.22	0.21	0.29	-0.02	0.04	-0.19
		300	700	0.40	0.40	0.42	0.02	0.07	-0.17
		300	400	0.39	0.38	0.42	-0.02	0.03	-0.22
		200	300	0.56	0.51	0.56	-0.02	0.05	-0.26
		150	300	0.59	0.53	0.57	-0.02	0.06	-0.24
		100	300	0.69	0.62	0.65	0.04	0.09	-0.21
		75	300	0.76	0.68	0.70	-0.03	0.05	-0.25
		50	300	0.94	0.89	0.91	-0.02	0.02	-0.30
57	1.63	700	2,100	0.23	0.23	0.32	-0.01	-0.01	-0.23
		300	700	0.46	0.40	0.50	-0.03	-0.05	-0.29
		300	400	0.49	0.42	0.51	0.06	0.03	-0.24
		200	300	0.60	0.53	0.59	0.01	0.03	-0.26
		150	300	0.61	0.52	0.62	-0.03	-0.03	-0.35
		100	300	0.78	0.57	0.69	0.00	0.01	-0.31
		75	300	0.95	0.79	0.85	0.03	-0.08	-0.32
		50	300	1.16	1.00	1.08	-0.19	-0.08	-0.47
59	1.81	700	2,100	0.26	0.23	0.34	-0.01	0.02	-0.23
		300	700	0.50	0.45	0.51	0.09	0.07	-0.22
		300	400	0.56	0.49	0.53	0.07	0.06	-0.25
		200	300	0.51	0.49	0.54	0.02	0.10	-0.24
		150	300	0.68	0.53	0.61	0.07	0.11	-0.23
		100	300	0.68	0.58	0.65	0	0.08	-0.30
		75	300	0.93	0.78	0.85	0.05	0.03	-0.27
		50	300	1.06	0.88	0.89	0.02	0.17	-0.25

Note. MH = Mantel-Haenszel, MHKS = Mantel-Haenszel kernel-smoothed, MHLS = Mantel-Haenszel loglinear-smoothed, RMSD = root-mean square difference.

sample estimates were closer to the criterion. In general, RMSD was smaller for large sample size combinations than for small ones, regardless of smoothing. This was expected, as estimates from larger samples should be closer to the criterion than those from the small samples. Compared to the

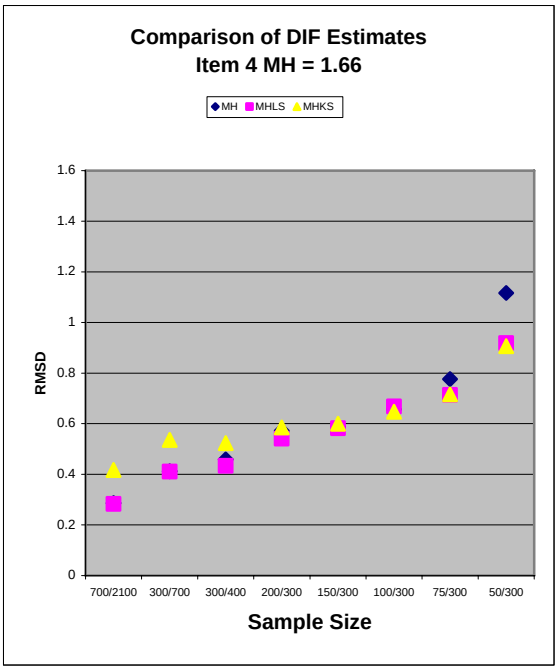
unsmoothed data, loglinear smoothing produced comparable or improved estimates. RMSD was .09 to .28 lower for MHLS than for MH for all the DIF items at the smallest sample size combination of 50/300. This degree of improvement in estimates can be seen at larger sample size combinations, up to 150/300 for Items 6, 7, and 59 and to 100/300 for Items 10 and 57. Item 49 is not a DIF item using ETS classification rules. It was included for the purpose of comparison. The difference in RMSDs between MH and MHLS for this item is not prominent.

Presented in Figures 11, 12, and 13 are plots of RMSDs for the three types of data across all sample size combinations for each item. In addition to demonstrating an increase in RMSDs with decreasing sample size combinations, the plots show that RMSDs for the unsmoothed data were larger than those for the smoothed ones at smaller sample size combinations. Noticeable differences can be observed at the sample size condition of 100/300 and below for most items (e.g., Items 6, 7, 10, 13, and 57). Another pattern to note is that RMSDs are the smallest at the largest sample size combination (700/2100) and are very similar at 300/700 and 300/400. They started to increase visibly from 200/300 for all the items except for Item 7, for which an increase commenced at 100/300.

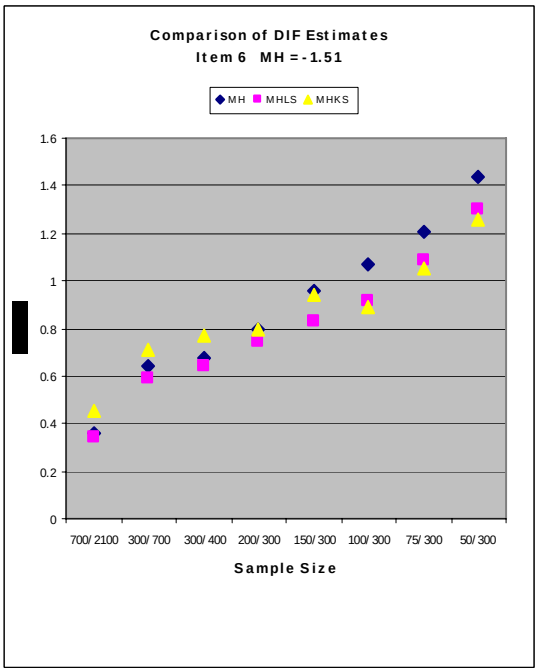
Bias analyses were also conducted, and the results are summarized in Table 3. In general, bias is similar for MH and MHLS across the sample size conditions. At the sample size combination of 50/300, the bias estimate is lower for MHLS than for MH by at least 1.0 for Item 4, but is higher for MHLS than for MH by at least 1.0 for items 7, 57, and 59. No method demonstrates a clear advantage. Bias estimates for MHKS, however, are much bigger compared to those for MH and MHLS and are negative for all items across all sample size combinations. This indicates that the samples consistently underestimate population DIF.

Type II Error Rate

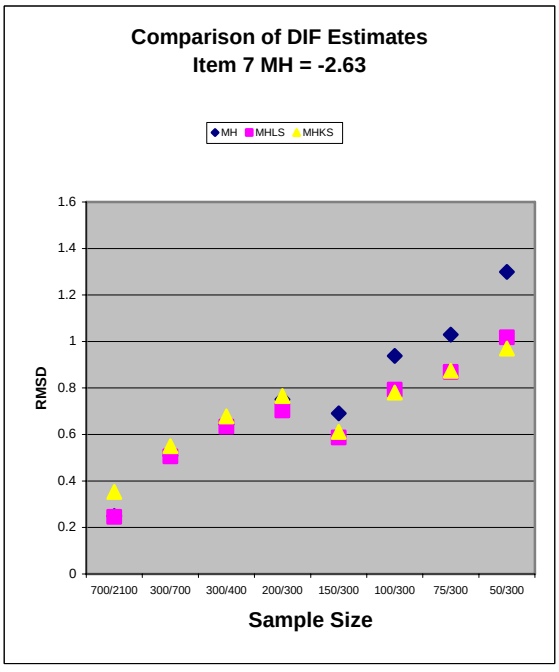
Type II errors occur when DIF that exists in the population is not detected in a particular method and sample size combination. The Type II error rate for each of the eight DIF items and each type of data was computed using the number of times that the item was not detected as a C DIF item divided by the 80 replications. The results for the eight DIF items are summarized in Table 4 and plotted in Figure 14. The smaller the numbers, the lower the Type II error rates. Lower error rates indicate more power.



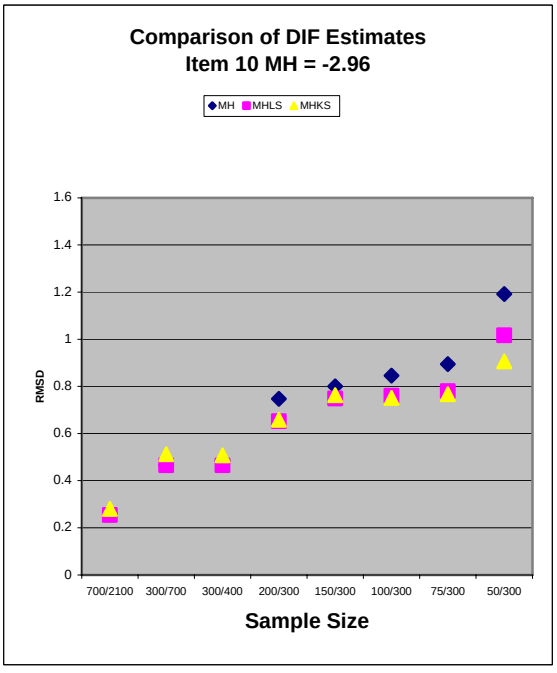
(a)



(b)

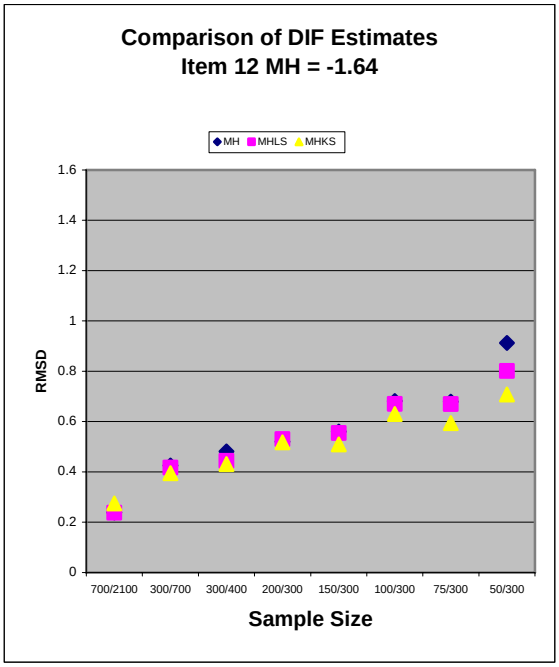


(c)

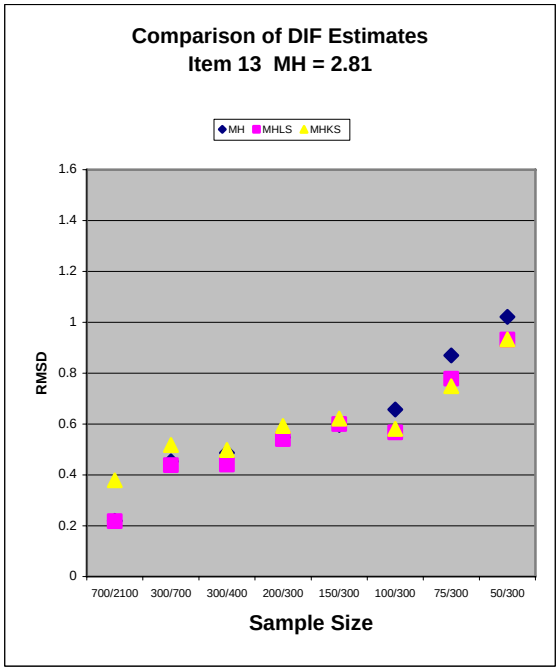


(d)

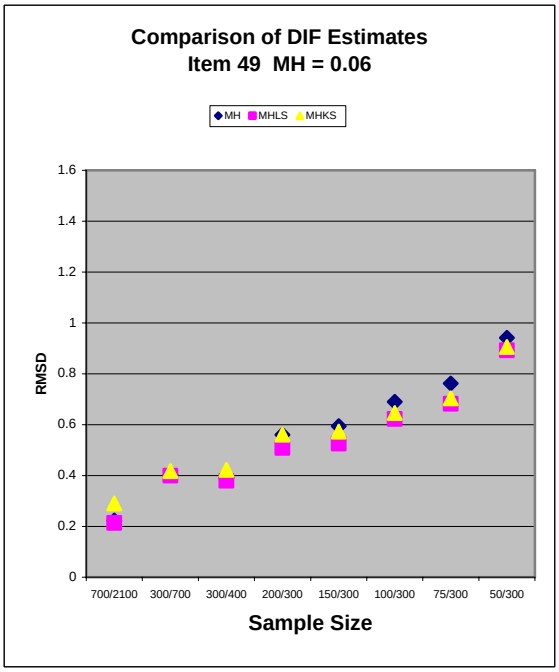
***Figure 11*. Plots of root-mean square difference for Items 4, 6, 7, and 10.**



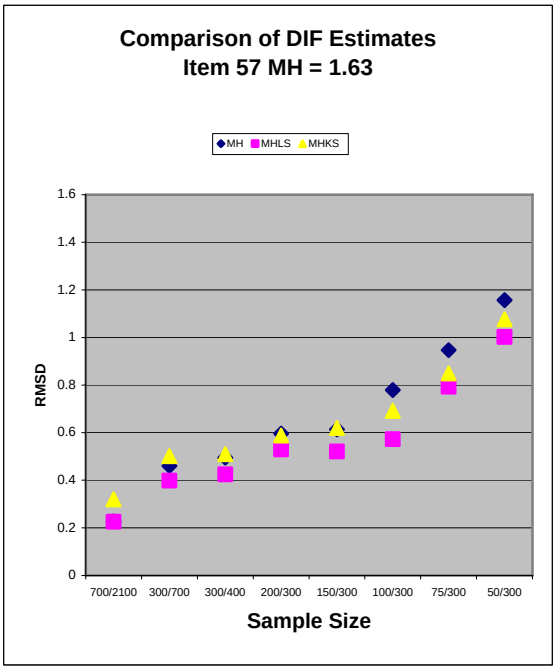
(e)



(f)



(g)



(h)

***Figure 12*. Plots of root-mean square difference for Items 12, 13, 49, and 57.**

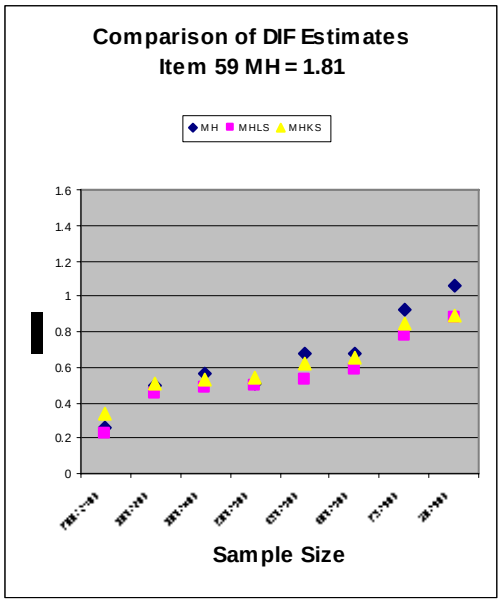


***Figure 13*. Plot of root-mean square difference for Item 59.**

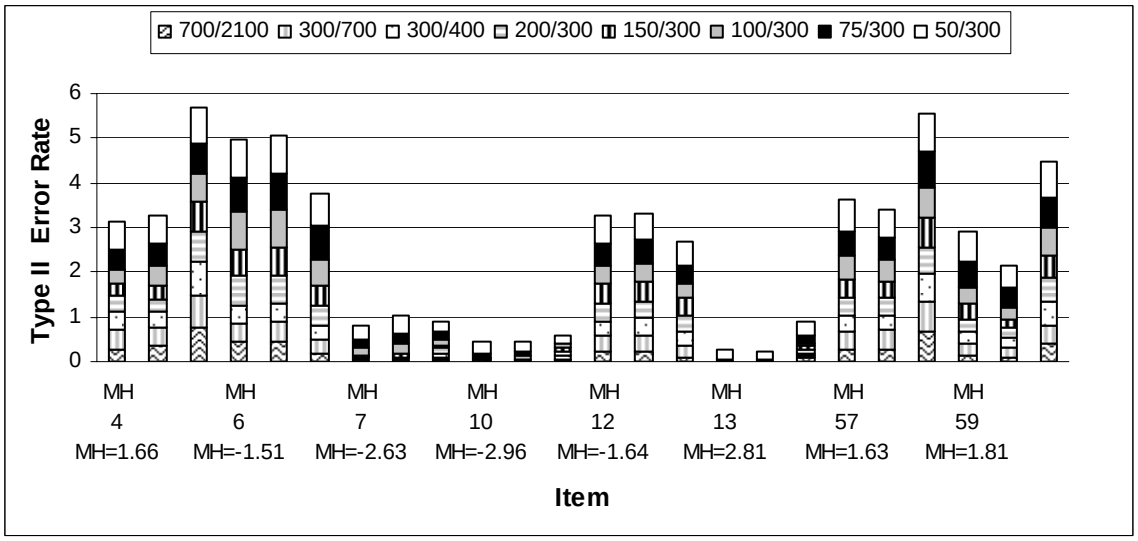


***Figure 14*. Bar graph for error rates for all items**

Mostly, Type II error rates increase when sample sizes become smaller, no matter whether smoothing is applied or not. For instance, the Type II error rate for Item 59 was .15 at 700/2100, .29 at 300/400, .36 at 150/300, and .56 at 75/300 for MH. This means that the item was not identified as containing C DIF 15% of the time at 700/2100, but 56% of the time at

Table 4

Type II Error Rates for DIF Items

700/2100	MH	MHLS	MHKS	300/700	MH	MHLS	MHKS
4	0.29	0.35	0.76	4	0.41	0.40	0.71
6	0.46	0.46	0.19	6	0.39	0.42	0.29
7	0.00	0.00	0.03	7	0.00	0.00	0.05
10	0.00	0.00	0.03	10	0.00	0.00	0.04
12	0.23	0.24	0.10	12	0.34	0.36	0.28
13	0.00	0.00	0.08	13	0.00	0.00	0.06
57	0.28	0.29	0.69	57	0.41	0.41	0.66
59	0.15	0.10	0.39	59	0.24	0.20	0.41
300/400	MH	MHLS	MHKS	200/300	MH	MHLS	MHKS
4	0.40	0.36	0.75	4	0.36	0.28	0.70
6	0.40	0.41	0.34	6	0.66	0.64	0.44
7	0.03	0.04	0.13	7	0.05	0.07	0.11
10	0.00	0.00	0.09	10	0.01	0.03	0.08
12	0.33	0.39	0.29	12	0.41	0.38	0.36
13	0.00	0.01	0.06	13	0.00	0.00	0.08
57	0.36	0.35	0.61	57	0.40	0.38	0.60
59	0.29	0.23	0.53	59	0.26	0.23	0.56
150/300	MH	MHLS	MHKS	100/300	MH	MHLS	MHKS
4	0.29	0.33	0.65	4	0.31	0.43	0.63
6	0.61	0.61	0.44	6	0.81	0.85	0.61
7	0.04	0.07	0.06	7	0.20	0.22	0.11
10	0.04	0.04	0.08	10	0.05	0.06	0.09
12	0.45	0.41	0.40	12	0.39	0.44	0.33
13	0.00	0.00	0.09	13	0.00	0.00	0.05
57	0.40	0.38	0.65	57	0.51	0.48	0.70
59	0.36	0.21	0.48	59	0.36	0.25	0.64
'5/300	MH	MHLS	MHKS	50/300	MH	MHLS	MHKS
4	0.43	0.48	0.70	4	0.64	0.65	0.79
6	0.78	0.81	0.73	6	0.84	0.86	0.75
7	0.20	0.24	0.16	7	0.29	0.39	0.23
10	0.09	0.10	0.03	10	0.26	0.22	0.16
12	0.50	0.50	0.39	12	0.63	0.59	0.54
13	0.04	0.05	0.15	13	0.21	0.17	0.33
57	0.55	0.50	0.78	57	0.73	0.63	0.85
59	0.56	0.44	0.69	59	0.70	0.50	0.81

Note. MH = Mantel-Haenszel, MHKS = Mantel-Haenszel kernel-smoothed, MHLS = Mantel-Haenszel loglinear-smoothed.

75/300. The error rate almost quadrupled when the sample size was cut down by nearly nine times.

Type II error rates are directly related to the amount of true DIF an item contains. The less the amount of true DIF, i.e., the closer the true DIF to the absolute value of 1.5, the higher the error rate across all sample size combinations. The error rates for all the eight DIF items are plotted in Figure 14. Higher error rates, symbolized by higher bars, can be clearly noted for Items 4, 6, 12, and 57, all of which have absolute criterion DIF values between 1.51 and 1.66, and for Item 59, which has a criterion DIF value of 1.81. On the other hand, much lower error rates can be easily observed for items 7, 10, and 13, which have absolute criterion DIF values between 2.63 and 2.81.

In general, error rates were similar for MH and MHLS, although they varied slightly for some items under some conditions. Results for MHKS demonstrate a different pattern compared to those for MH and MHLS. The error rates for MHKS are inconsistent with their counterparts for MH and MHLS, being much larger (as for Items 4, 13, 57, and 59) across most sample size combinations, or much smaller (as for Item 6).

Summary and Discussion

This study investigated DIF estimation with realistic small samples by applying the MH procedure to the unsmoothed or raw data, loglinear-smoothed data, and kernel- smoothed data. The results show that sample estimates of population DIF are more variable at smaller sample sizes than at larger ones, as expected. The MH estimates as summarized in boxplots were very similar for the unsmoothed and loglinear-smoothed data, with loglinear smoothed estimates less variable, especially at smaller sample size combinations of 75/300 and 50/300 for most items. Estimates produced from kernel smoothed data showed a pattern different from those for the other two. The center of the boxplot deviated from the criterion for all items across all sample size conditions.

RMSDs between the average sample estimate over 80 replications and the criterion values increased when sample sizes became smaller, regardless of smoothing. Loglinear smoothing produced comparable or slightly improved estimates when compared to those without smoothing. RMSDs were smaller for the smallest sample size combinations for all DIF items when smoothing was applied.

Bias estimates from the unsmoothed and loglinear-smoothed data are very similar. At the smallest sample size combinations, bias was smaller for unsmoothed estimates than for loglinear-smoothed estimates for some items but bigger for other items. With kernel smoothing, however, bias tended to be considerably larger than that for both unsmoothed and loglinear-smoothed estimates and consistently underestimated population DIF no matter what sample size combination was used.

Type II error rates also increase when sample sizes decrease. And they are related to the amount of true DIF an item contains. When true DIF is larger, it is more readily identified in the samples, thereby producing lower error rates. The error rates for unsmoothed and loglinear-smoothed estimates are very similar, while those for kernel-smoothed estimates have their own pattern and are not related to those for the other two.

The unsmoothed and loglinear-smoothed results are similar in terms of bias and Type II error rates. Loglinear smoothing demonstrated moderate improvements in DIF estimation with small samples in that, at the smallest sample size conditions, it produced sample estimates that were less variable and RMSDs that were smaller. Smaller RMSD indicates that sample estimates are closer to the true DIF. The reduction in RMSD for loglinear-smoothed estimates could produce comparably more accurate DIF estimation at sample size combinations less than 200/300. These results are encouraging, although more research is needed to replicate these findings before applying the MH procedure to the loglinear smoothed data in operational settings.

On the other hand, kernel smoothing produced results different from those from unsmoothing and loglinear smoothing. In general, sample estimates from kernel smoothing deviated from the criteria obtained from the population, thereby producing large biased estimates for all items across all sample size conditions. And all the biased estimates were negative, indicating the under-estimation of population DIF. The error rates from kernel smoothing were also somewhat irregular and inconsistent when compared to those from unsmoothing and loglinear smoothing. These results suggest that applying MH to kernel smoothed data, as implemented in this study, is not an appropriate option for DIF estimation in small samples. In an exploratory analysis (Yu, Moses, Puhan, & Dorans, 2005), when MH was applied to kernel-smoothed odds ratios, the results were similar to those from MH using unsmoothed or raw data. In summary, kernel smoothing, as implemented here, does not seem to improve DIF estimations.

The source of the kernel smoothed biases seems to be the result of smoothing that was too strong or the result of the bandwidths that were too large. A different selection of the kernel bandwidths than the selection rule of $1.1N^{-.2}$ may have produced less biased but more varying kernel results. This bandwidth rule is the default ETS rule for smoothing item probability plots (Ramsay, 1991) and likely was not appropriate for frequency distributions. In comparison to the loglinear smoothing, which used one model across the sample size combinations, the selection rule in kernel smoothing that applied more smoothing to small samples automatically biased results for the smallest samples.

Sample MH statistics based on loglinear smoothing were similar to the population MH values, and the sample MH statistics based on kernel smoothing were not similar when compared to the population MH values, because both the MH and loglinear smoothing are based on similar methodologies—the loglinear models. The MH uses loglinear models based on a noniterative fitting process, while the loglinear smoothing uses similar, but not identical, loglinear models based on an iterative fitting process. These two methods, therefore, produce results that are expected to be similar. On the other hand, kernel smoothing is based on a very different methodology—computing moving averages as defined by the kernel bandwidth. The criterion used to evaluate the results is obtained in the population using the MH. Consequently, results from kernel smoothing tend to be worse, if they differ from those of the MH and loglinear smoothing at all. The selection of the bandwidth in kernel smoothing and the moment-matching in loglinear smoothing are fundamentally different, which makes it less likely for kernel smoothing to work as well as or as closely as loglinear smoothing.

The results of the current study demonstrate greater power in MH DIF detection than was found by Mazor et al. (1992). In their study using simulated data that contained DIF items, the MH procedure failed to detect 30% of the DIF items at sample size of 2,000 and 50% of them at sample size of 500 or fewer. This study shows that the rate at which DIF was not detected for 7 of the 8 items was equal to or less than 29% at the focal group sample size of 700 and 45% at the focal group sample size of 150. When loglinear smoothing was applied, the error rate was less than or equal to 44% for these seven items at the focal group sample size of 100. The error rates for Item 6 were large across all sample size combinations, because the population DIF value of -.1.51 was too close to the criterion used to flag significant DIF.

Loglinear smoothing provides improvements in MH DIF estimation with small samples. In general, loglinear smoothing worked well in the procedure. Nonconvergence, however, does exist. Nonconvergence is a numerical problem encountered when a solution cannot be reached within a given level of tolerance after a given number of iterations. It is usually caused by the number of available cases in a cell. When samples are small, the number of cases in each of the distributions that smoothing applies to get even smaller. In case of nonconvergence in the current study, results were analyzed only for the converging samples. There were usually 5 or 6 or fewer cases that failed to converge. Alternative ways of implementing the loglinear smoothing could improve the convergence rates. These include using simpler smoothing models, using orthogonal polynomials rather than power functions of the test scores, or using less stringent convergence criteria.

Future research may focus on replicating the results of this study. This research shows that DIF estimation in small samples improves slightly when data are smoothed using loglinear models. If kernel smoothing is used, a different band width should be attempted for unbiased estimates. Alternative strategies should be considered for selecting kernel bandwidths that are smaller and less likely to produce biased results in the smoothing of frequencies. It is also important to know if such a finding can be replicated on data sets other than admission tests. Future research may consider using other types of data, such as licensure exams and K-12 assessments. In licensure exams, small samples occur frequently. On many occasions, DIF cannot be performed because of lack of sufficient data. In recent years, large-scale state assessments are growing rapidly. Although the population of test-takers may be huge, some minority groups (such as Native Americans or Pacific Islanders) may still have small numbers for performing DIF analysis. Conducting DIF on small samples is important so that the items in a test measure what they are supposed to measure regardless of the gender, the ethnic group, or the special education status of the examinees.

The data used in this study were formula scored, and the score distributions contained teeth at regular intervals. Future research may use rights-scored data and examine the effects of smoothing on such data in DIF estimation with small samples.

The test used in this study consisted of 60 items, with eight DIF items. Test length and the number of DIF items can also be varied in future studies to see if loglinear smoothing improves estimation. Of note, the study was also completed by applying the standardization

approach (Dorans & Kulick, 1986) to the data, using the same smoothing techniques and using the same items and same sample size conditions. The results were similar to those for the MH method.

References

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing the unexpected differential item functioning on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355–368.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (ETS Research Rep. No. RR-87-31). Princeton, NJ: ETS.

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

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25*(2), 133–183.

Lyu, C. F., Dorans, N. J., & Ramsey, J. O. (1995). *Smoothed standardization assessment of testlet level DIF on a math free-response item type* (ETS Research Rep. No. RR-95-38). Princeton, NJ: ETS.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52,* 443–451.

Parshall, C. G., & Miller, T. R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small-sample conditions. *Journal of Educational Measurement, 32*(3), 302–316.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611–630.

Roussos, L. A., & Stout, W. F. (1996). 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*(2), 215–230.

Sinharay, S., Dorans, N. J., Grant, M. C., Blew, E. O., & Knorr, C. M. (2006). *Using past data to enhance small-sample DIF estimation: A Bayesian approach* (ETS Research Rep. No. RR-06-09). Princeton, NJ: ETS.

Yu, L., Moses, T., Puhan, G., & Dorans, N. (2005, April). *Differential item functioning estimation with small samples*. Paper presented at the annual meeting of the California Educational Research Association, Long Beach, CA.

Zwick, R., & Thayer, D. T. (2002). Application of an empirical Bayes enhancement of Mantel-Haenszel DIF analysis to a computerized adaptive test. *Applied Psychological Measurement, 26*(1), 57-76.

Zwick, R., Thayer, D. T., & Lewis, C. (1997). *An investigation of the validity of an empirical Bayes approach to Mantel-Haenszel DIF analysis* (ETS Research Rep. No. RR-97-21). Princeton, NJ: ETS.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenzel DIF analysis. *Journal of Educational Measurement, 36*(1), 1–28.

Zwick, R., Thayer, D. T., & Lewis, C. (2000).Using loss function for DIF detection: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics, 25*(2), 225–247.