DOCUMENT RESUME

ED 445 011 TM 031 548

AUTHOR	Chu, Kwang-lee; Kamata, Akihito
TITLE	Nonequivalent Group Equating via 1-P HGLLM.
PUB DATE	2000-04-00
NOTE	35p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Equated Scores; *Groups; Models
IDENTIFIERS	Hierarchical Analysis; One Parameter Model

ABSTRACT

The quality of nonequivalent group equating by the one-parameter hierarchical generalized linear logistic model (1-P HGLLM) was examined by comparing it with: (1) traditional concurrent equating; (2) Stocking-Lord's method; and (3) multiple-group concurrent equating. Root mean squared errors (RMSEs) for item parameters indicated that there was no prominent difference among the four equating methods, and none of the four methods was consistently better than other methods across the entire item difficulty range. RMSEs for ability parameters of 1-P HGLLM were similar to those of traditional concurrent equating, which resulted in higher RMSEs than the Stocking-Lord method and multiple-group concurrent equating. The 1-P HGLLM method did not show advantages compared to other equating methods, but it did not show many disadvantages either. It is suggested that the equating model be extended in situations in which the effects of persons and group characteristics on performance are of interest. (Contains 6 tables, 18 figures, and 13 references.) (Author/SLD)

Nonequivalent Group Equating Via 1-P HGLLM

Kwang-lee Chu

Akihito Kamata

Florida State University

March 28, 2000

CORRESPONDING AUTHOR: Kwang-lee Chu
307 Stone Building
Department of Educational Research
Tallahassee, FL 32306-4453
(850) 644-4592
klchu@garnet.acns.fsu.edu

Paper presented at the annual meeting of the American Educational Research Association, April, 2000; New Orleans.

Abstract

The quality of nonequivalent group equating by 1-P HGLLM is examined by comparing with (a) traditional concurrent equating, (b) Stocking-Lord's method, and (c) multiple-group concurrent equating method. Root mean squared errors (RMSEs) for item parameters indicated that there was no prominent difference among the four equating methods and none of the four methods was constantly better than the other methods across the entire item difficulty range. RMSEs for ability parameters of 1-P HGLLM were similar to the traditional concurrent equating, which resulted in higher RMSEs than Stocking-Lord's methods and multiple-group concurrent equating. 1-P HGLLM did not show advantages compared to other equating methods, while it did not show many disadvantages either. It is suggested that the equating model to be extended in situations where the effects of person and group characteristics on performance are of interested.

Nonequivalent Group Equating Via 1-P HGLLM

Multiple-group item response theory (MG-IRT) has been developed in the context of estimating group-level abilities in multiple matrix sampling (e.g., Bock & Mislevy, 1981; Mislevy, 1983). Recent presentations (e.g., Bock & Zimowski, 1996) has clarified that MG-IRT can be applied into many other settings, including nonequivalent-group equating. Since MG-IRT assumes separate latent distributions for separate groups when item parameters are estimated, MG-IRT theoretically fits nicely with nonequivalent-group equating.

Recent studies (e.g., Hedges & Vevea, 1997; Kim & Cohen, 1998; and Hanson & Beguin, 1999) compared the performance of nonequivalent-group equating by MG-IRT concurrent equating with traditional equating methods, such as traditional concurrent equating and Stocking-Lord procedure. These studied showed that the results depended on the assumptions made in the models. Procedures that assume different means, standard deviations, and shapes for separate latent distributions consistently showed more satisfactory outcomes than procedures with more restrictive assumptions.

Kamata (1998) also proposed a multiple-group model. He demonstrated that the Rasch model can be formulated as a special case of hierarchical generalized linear model (HGLM) (Raudenbush, 1995). The reformulated Rasch model is referred to as one-parameter hierarchical generalized linear logistic model (1-P HGLLM). He referenced several extensions of 1-P HGLLM, including a multilevel item response model. This particular extension can be applied, but are not limited to nonequivalent group concurrent equating. However, there is no study that investigated the quality of nonequivalent group concurrent equating by 1-P HGLLM.

The purpose of this study is to investigate the performance of 1-P HGLLM nonequivalent group equating quality. The quality of the equating is to be compared with (a) traditional concurrent equating, (b) Stocking-Lord's method, and (c) MG-IRT concurrent equating method.

1-P HGLLM as a Concurrent Equating Model

For item i ($i = 1, \ldots, k$) and person j ($j = 1, \ldots, n$) in group m ($m = 1, \ldots, r$), the level-1 structural model is defined as

$$\begin{aligned}\log\left(\frac{p_{ijm}}{1-p_{ijm}}\right) &= \eta_{ijm} \\ &= \beta_{0jm} + \beta_{1jm}X_{1jm} + \beta_{2jm}X_{2jm} + \cdots + \beta_{(k-1)jm}X_{(k-1)jm} \\ &= \beta_{0jm} + \sum_{i=1}^{k-1}\beta_{ijm}X_{ijm}\ ,\end{aligned} \tag{1}$$

where X_{ijm} is the ith dummy variable for person j in group m, with a value of 1 when the observation is the ith item, and 0 otherwise. The coefficient β_{0jm} is an intercept term, and β_{ijm} is a coefficient associated with X_{ijm}, where $i = 1, \ldots, k-1$. Here, the model assumes the coefficient for the last item to be constrained as 0. The model can be reduced to

$$\log\left(\frac{p_{ijm}}{1-p_{ijm}}\right) = \eta_{ijm} = \beta_{0jm} + \beta_{ijm}\ , \tag{2}$$

for item i, given $X_{ijm} = 1$ for the ith item and 0 otherwise. This way, β_{ijm} represents the effect of the ith item. Here, β_{0jm} is an intercept term and is considered to be an overall effect common to all items, in effect, the mean effect of items, with the constraint $\beta_{kjm} = 0$. On the other hand, β_{ijm} represents the specific effect of the ith item for $i = 1, \ldots, k-1$. Then the probability that person j in group m answers item i correctly is expressed as

$$p_{ijm} = \frac{1}{1+\exp\left[-\eta_{ijm}\right]}, \tag{3}$$

which follows from Equation 2.

The level-2 models are person-level models, which specify that item effects are constant across people. Therefore, the level-2 models are

$$\begin{cases} \beta_{0jm} = \gamma_{00m} + u_{0jm} \\ \beta_{1jm} = \gamma_{10m} \\ \quad\vdots \\ \beta_{(k-1)jm} = \gamma_{(k-1)0m} \end{cases}, \tag{4}$$

where u_{0jm} is a random component of β_{0jm} and distributed as $N(r_{00m},\ \tau_\gamma)$, which states that u_{0jm} is normally distributed with the mean of r_{00m}. Also, the variance of u_{0jm} within the group is denoted τ_γ and is assumed to be identical for all groups.

Now, the level-3 model, a school-level model, could show that item effects are constant across schools. The overall effect of items, γ_{00m}, is the only term that varies across schools. For school m, we have

$$\begin{cases} \gamma_{00m} = \pi_{000} + r_{00m} \\ \gamma_{10m} = \pi_{100} \\ \gamma_{20m} = \pi_{200} \\ \vdots \\ \gamma_{(k-1)0m} = \pi_{(k-1)00} \end{cases}, \tag{5}$$

where $r_{00m} \sim N(0, \tau_\pi)$. Here π_{000} is the fixed component of γ_{00m}, r_{00m} is the random component of γ_{00m}, and τ_π is the variance of r_{00m}. On the other hand, γ_{10m} through $\gamma_{(k-1)0m}$ have only fixed components, i.e., π_{100} through $\pi_{(k-1)00}$. As a result, the combined model is expressed as

$$p_{ijm} = \frac{1}{1+\exp\left\{-\left[-(r_{00m}+u_{0jm})-(\pi_{i00}+\pi_{000})\right]\right\}}. \tag{6}$$

This parallels the Rasch model, where $-(r_{00m} + u_{0jm})$ is the ability of person j in group m, and $-(\pi_{i00} + \pi_{000})$ is the difficulty of item i. The abilities for this three-level model consist of two parts. First, r_{00m} is the random effect associated with school m, and can be interpreted as the average ability of students in school m. Second, u_{0jm} is a person-specific ability of person j in school m, indicating how much the ability of person j is deviated from the average ability of the students in school m. This way, the three-level model can provide school abilities, as well as individual person abilities.

This formulation allows for missing data, which still being able to estimate parameters. In other words, examinees do not have to respond all the items. Therefore, the above mentioned model can be directly applied to concurrent equating of test items from more than one test form, where we assume that a sample of examinees take one of the test forms. When we have common items between test forms, item parameter estimates across forms are estimated on the same scale.

Another characteristic of this hierarchical model is that all of the sub-populations have the same shape of latent distributions. In other words, the standard deviation all of the sub-populations are the same (homogeneity of variances) and normally distributed, although the means could be different. This assumption is embedded in equation 4 and 5, where π_{000} is the mean of sub-population means and γ_{00m} is the sub-population mean for group m. The standard deviation of the sub-populations is τ_{γ}, and is identical for all groups.

Other Equating Procedures

1-P HGLLM equating results are compared with (a) traditional concurrent equating, (b) Stocking-Lord's method (Stocking & Lord, 1983), and (c) MG-IRT concurrent equating method. Traditional concurrent equating is a one-step equating procedure, which does not require a separate step to put item and person parameters on a common scale. It assumes samples are from one underlying latent population. Then, it uses the information of combined latent distribution, rather than using the information of possibly different sub-populations separately, when item parameters are estimated.

Since Stocking-Lord's method (S-L) calibrates item parameters separately for each group, it automatically assumes the sub-populations can have different distribution characteristics. S-L is a two-step equating procedure, where the first step is to estimate parameters from different test forms, and the second step is to equate parameters of different test forms onto a common scale using characteristic curve transformation method.

Like the traditional concurrent equating procedure, MG-IRT concurrent equating method is a one-step equating procedure. It assumes an underlying normally distributed latent population, then uses the characteristics of latent distributions separately for each of the sub-populations during item parameter estimations. This allows sub-populations to have different distributions, that is, different means, standard deviations, and shapes.

As described above, each equating method assumes different assumptions for latent distributions of groups. 1-P HGLLM assumes latent distributions are all normal with the same standard deviations, but different means. The traditional concurrent equating only assumes the mean and standard deviation of the combined latent distribution, where the shape of the distribution can be freely estimated. Stocking-Lord method and MG-IRT assume separate latent distributions for groups with no restrictions for the means, standard deviations, and the shapes of the distributions.

It is reasonable to expect that when its assumptions are met, 1-P HGLLM should have compatible equating results to MG-IRT concurrent equating and Stocking-Lord procedure. Also, it is expected that 1-P HGLLM performs better than or equally well as the traditional concurrent equating, unless the standard deviations of the latent distributions are extremely different between groups and/or the shape of the latent distributions is extremely different from normal.

Methods

As mentioned above, a 3-level 1-P HGLLM was employed to conduct a non-equivalent group concurrent equating. The performance of the equating by 1-P HGLLM was then compared to the three other equating methods mentioned above.

It was assumed that two tests were given to two separate samples and each test contained 20 items, including 5 common items. Item difficulties were arbitrarily chosen, so that those in Form X ranged from – 2.3 to 2.5, and those in Form Y ranged from – 2.2 to 2.6. The values of item difficulties are listed in Table 1. Item difficulties for the 5 common items were arbitrarily chosen to be – 0.7, – 0.6, 0.1, 0.7, and 0.9.

True ability values were generated so that they were distributed normally for each sample. 200 examinees were assumed in each sample. Also, it was assumed that the ability distribution for the second group (group B) had higher mean and/or smaller standard deviation than the first group (group A). The ability distribution for group A had the mean of 0 and the standard deviation of 1 for all conditions. On the other hand, for group B, the mean was one of 0, 0.5, or 1.0, and the standard deviation was one of 1.0, 0.75, or 0.5. As a result, 9 different conditions of the ability distribution for group B were created, and equating was performed between Form X, taken by group A and Form B, taken by group B with one of 9 distribution conditions. The conditions of equating were summarized in Table 2. Equating was replicated 20 times for each one of the 9 equating conditions for each method.

1-P HGLLM estimation procedure was conducted by HLM (Bryker, Randembush, & Congdon, 1996). Traditional concurrent equating was conducted using BILOG (Mislevy & Bock, 1990) for both parameter estimates and equating. Stocking-Lord's method used BILOG for group parameter estimates and ST (Hanson & Zeng, 1995) for calibration. MG- concurrent equating method used BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1999).

In order to assess the quality of equating, room mean squared error (RMSE) was calculated for all item and ability parameters. Also, the mean RMSE for items and abilities was computed for each equating condition as an index of overall equating performance.

Results

RMSE for Item Parameters

Item parameter RMSEs were similar through out all 9 conditions. In all of the 9 conditions, RMSEs increased when item difficulty moved toward extremes (see Figure 1 to 9). The RMSEs of common items were not smaller than non-common items with similar difficulties. Since common items had twice the sample size of non-common items, this finding indicated that 200 hundred samples were sufficient for item parameter estimates in this study.

When group means were equal (0 in this case) and standard deviations were different, small fluctuation of RMSEs were observed for all four methods (see Table 3). When the standard deviation of group B increased, the fluctuation of RMSEs became slightly larger. However, the majority of RMSE differences between conditions were smaller than 0.1 and only a few cases had differences larger than 0.5. By comparing Figures 1, 2, and 3, larger differences were found on ST for some items with low difficulty. At both extremes of item difficulty, the RMSEs of the four methods vacillated, but BILOG-MG tended to have higher RMSEs at the higher end.

When standard deviations were the same (1 in this case) and group mean difference increased, the absolute values of BILOG-MG's RMSE differences between conditions remained about the same, while other methods' increased (see Table 4). More than 70% of items of HLM and BILOG's RMSE differences were more than 0.1 regardless of the magnitude of the mean difference, while ST had more than 50% of items. When the means were 0 and 1 for group A and group B, respectively, RMSE differences of 3 items were more than 1 for HLM and BILOG, while 7 for ST.

When comparing Figures 1, 4, and 5, the RMSEs vacillated across the four methods. When standard deviations were different between groups, the RMSEs of HLM fluctuated for items with lower difficulties and were generally higher for items with higher difficulties. On the other hand, the RMSEs of ST vacillated more for items with higher difficulties and were generally less for items at the lower end.

When both means and standard deviations were different between groups, the pattern of RMSEs across conditions was more similar to the pattern when only means between groups were different. In other words, mean differences affected equating quality more than standard deviation differences did.

An investigation of the mean and standard deviation of RMSEs (see Table 5) revealed that there were patterns that coincided with mean differences between groups. HLM had the highest mean across the 9 conditions, while BILOG, ST, and BILOG-MG had the lowest mean when the mean of group B was 0, 0.5, and 1, respectively. However, most of the mean differences were at the second decimal point. When the mean of group B was 0, HLM had the lowest standard deviation and BILOG-MG had the highest. When the mean of group B was 0.5, ST had the lowest standard deviation and BILOG had the highest. When the mean of group B was 1, BILOG_MG had the lowest standard deviation and BILOG had the highest one. A standard deviation comparison between HLM and BILOG revealed that HLM had consistently smaller standard deviation than BILOG. This could be a result of the shrinkage of Empirical Bayes estimates.

The inspection of RMSEs for item parameters led to two conclusions. First, mean differences had more effect on parameter estimates than the differences in standard deviations, except for BILOG-MG. When means were different, BILOG-MG resulted in slightly more consistent estimation than standard deviations were different. Second, there was not a single method that performed consistently better than other methods across all item difficulty levels.

RMSE for Ability

The person parameter RMSEs for the group A were identical across conditions for all methods (see Figure 10 to 18). An investigation across the four methods revealed that HLM had similar RMSEs pattern to BILOG and ST. Their RMSEs distributed almost symmetrically around the mean of theta (0 in this case). As the theta value moved away from the mean, higher RMSEs and larger dispersions were observed. When comparing their RMSE values, HLM had higher RMSEs than the other methods. BILOG and ST had similar RMSEs. On the other hand, BILOG-MG RMSEs for the group A clustered tightly across the theta scale with larger RMSEs and dispersions at both ends.

In the comparisons of the person parameter RMSEs for the group B when group means were equal and group B standard deviation decreased, RMSEs dropped across the four methods. This could be a result of group homogeneity. Investigation of RMSE patterns across methods revealed that HLM was similar to BILOG, while BILOG-MG and ST were similar to each other. Higher RMSEs and larger dispersions were found in HLM and BILOG. Both BILOG-MG and ST showed a curved line of the RMSEs for group B with higher values at both ends. The line indicated a consistent estimation of group B person parameters.

When standard deviations were fixed and group B mean increased, all of the four methods had higher RMSEs at the higher theta end and lower RMSEs at the lower end. This could be because the common item difficulty range was out of group B's ability range. When group B mean was 0.5 or 1, the common item difficulties (ranges from –0.7 to 0.9) were at the lower end of the group B ability distribution. Hence, better estimation occurred at the lower end and more errors at the higher end.

Regardless of the common item difficulties range problem, RMSEs from BILOG-MG for both the group A and B lined up as a curved line when group means were different, which was not observed when group means were the same.

Inspection of the mean and standard deviation of RMSEs (see Table 6) across the 9 conditions revealed that the mean and standard deviation of RMSEs decreased as group B standard deviation decreased when group means were fixed. On the other hand, when group standard deviations were fixed, both mean and standard deviation of RMSEs increased as group B mean increased. When both means and standard deviations were different between groups, the mean differences had more impact on RMSEs than the standard deviation difference. When comparing the mean and standard deviation of RMSEs across the four methods, BILOG-MG and ST were smaller than HLM and BILOG. The differences were around 0.5 for both.

Three conclusions could be drawn from the investigations of person parameter RMSEs. First, both mean and standard deviation differences between groups affected person parameter RMSEs. Second, mean differences had higher impact on RMSEs than standard deviation differences. Third, when the magnitude of mean differences increased, its impact increased as well.

Summary and Discussions

Investigations of RMSEs for item parameters indicated that there was no prominent differences among the four equating methods and none of the four methods was constantly better than the other methods across the entire difficulty range. Although 1-P HGLLM had higher mean RMSE of item parameters than the other methods, the differences were less than 0.1 for all equating conditions. On the other hand, RMSEs for ability parameters were generally smaller for multiple-group concurrent equating. Person parameter estimates from multiple-group concurrent equating was much more stable than the other 3 methods, especially when two groups had different means and standard deviations. Throughout the 9 conditions, 1-P HGLLM results were very similar to traditional concurrent equating.

It was disappointing that 1-P HGLLM did not show its expected strengths. It was expected that 1-P HGLLM would show comparable results to MG-IRT and Stocking-Lord procedure, especially when two groups had the same standard deviations but different means. Instead, the results from 1-P HGLLM were more similar to the traditional concurrent equating. Therefore, we conclude that the use of 1-P HGLLM for the purpose of equating does not provide any advantage to other equating methods.

However, at the same time, it was not a disadvantage to use 1-P HGLLM in non-equivalent group equating either, because it performed as well as the traditional concurrent equating method. This encourages us to further extend the model to a situation where one is interested in investigating the effects of person- and group-level characteristics variables on the performance on tests. In cases such as examinees take different forms of a test, and examinees take different tests year-to-year, the comparisons of scores have to be based on equated scores. By including person- and group-characteristic variables in the 1-P HGLLM equating model, it achieves a 3-in-1 model, where scoring, equating, and analyses of person- and group-characteristic variables are performed in one step. This type of extension is currently possible only by 1-P HGLLM, and it is an obvious next step to conduct a real data analysis to answer real research question using such a model. Also, one shortcoming of this study was that common item difficulties were out of the ability range of group B in some conditions, which might have resulted in unconditionally unstable estimation of parameters.

References

Bock, R. D., & Mislevy, R. J. (1981). An item response model for matrix-sampling data: the California grade-three assessment. *New Directions for Testing and Measurement*, 10, 65-90.

Bock, R. D., & Zimowski, M. (1989). *Duplex Design: Giving Students a Stake in Educational Assessment*. Methodology Research Center, NORC. Chicago. IL

Bryk, A., Raudenbush, S., & Congdon, R. (1996). *Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs*. Scientific Software International, Inc. Chicago, IL.

Hanson, B., & Zeng, L. (1995). *ST, a computer program for IRT sclae transformaiton version 1.0*. ACT. Iowa City, IA.

Hanson, B. A., & Beguin, A. A. (1999). *Obtaining a common scale for IRT item parameters using separate versus concurrent estimation in the common item nonequivalent groups equating design.* Paper presented at the annual meeting of the American Educational Research Association, April, Montreal, Canada.

Hedges, L. V., & Vevea, J. L. (1997). A study of equating in NAEP. *NAEP Validity Studies*. Palo Alto, CA.

Kamata, A. (1998). *Some generlaizations of the Rasch model: an application of the hierarchical generalized linear model*. Unpublished doctoral dissertation. Michigan State University.

Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement,* 22, 131-143.

Mislevy, R. J. (1983). Item response models for grouped data. *Journal of Educational Statistics*, 8, 271-288.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3*. Scientific Software International, Inc. Mooresville, IL.

Raudenbush, S. W. (1995). *Posterior modal estimation for hierarchical generalized linear models with application to dichotomous and count data.* Unpublished manuscript. Michgan State University.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: multiple-group IRT analysis and test maintenance for binary items*. Scientific Software International, Inc. Chicago, IL.

Table 1. Item Difficulty of Form 1 and Form 2

Item Difficulty	Form 1	Form 2
Common item 1	-0.7	
Common item 2	-0.6	
Common item 3	0.1	
Common item 4	0.7	
Common item 5	0.9	
Item 1	-2.3	-2.2
Item 2	-1.9	-1.8
Item 3	-1.5	-1.4
Item 4	-1.1	-1.0
Item 5	-0.7	-0.6
Item 6	-0.2	-0.4
Item 7	-0.1	-0.2
Item 8	0.0	0.0
Item 9	0.2	0.2
Item 10	0.5	0.5
Item 11	0.9	1.2
Item 12	1.3	1.6
Item 13	1.7	2.2
Item 14	2.1	2.4
Item 15	2.5	2.6

Table 2. Sampling Distributions of Group A and B

Condition	Group A vs. Group B
Condition 1	N(0,1) vs. N(0,1)
Condition 2	N(0,1) vs. N(0, 0.75)
Condition 3	N(0,1) vs. N(0, 0.5)
Condition 4	N(0,1) vs. N(0.5, 1)
Condition 5	N(0,1) vs. N(0.5, 0.75)
Condition 6	N(0,1) vs. N(0.5, 0.5)
Condition 7	N(0,1) vs. N(1, 1)
Condition 8	N(0,1) vs. N(1, 0.75)
Condition 9	N(0,1) vs. N(1, 0.5)

Table3: RMSE Differences of Item Difficulty When Group Means are Fixed

	n(0,1) vs.n(0,1) - n(0,1) vs.(0,.75)				n(0,1) vs.n(0,.75) - n(0,1) vs.n(0,.5)				n(0,1) vs.n(0,1) - n(0,1) vs.(0,.5)			
b's	**HLM**	**BILOG**	**B-MG**	**ST**	**HLM**	**BILOG**	**B-MG**	**ST**	**HLM**	**BILOG**	**B-MG**	**ST**
-2.3	0.011	0.060	-0.001	0.000	0.127	0.050	0.000	0.000	0.138	0.110	-0.001	0.000
-2.2	-0.013	-0.230	-0.233	-0.280	0.115	-0.227	-0.263	-0.259	0.102	-0.457	-0.496	-0.539
-1.9	-0.015	0.003	-0.001	0.000	0.058	0.012	0.000	0.000	0.043	0.015	-0.001	0.000
-1.8	0.214	-0.117	-0.118	-0.147	0.301	-0.152	-0.177	-0.180	0.515	-0.268	-0.294	-0.327
-1.5	-0.013	0.013	-0.001	0.000	0.067	0.016	0.000	0.000	0.055	0.029	-0.001	0.000
-1.4	0.099	0.192	0.148	0.264	0.215	0.231	0.205	0.292	0.314	0.423	0.353	0.556
-1.1	-0.017	-0.003	0.000	0.000	0.039	0.001	0.000	0.000	0.023	-0.002	0.000	0.000
-1	-0.077	0.008	-0.007	0.023	-0.061	0.004	-0.013	0.012	-0.138	0.012	-0.021	0.035
-0.7	-0.004	0.011	0.006	0.000	-0.018	0.003	-0.004	0.000	-0.022	0.014	0.002	0.000
-0.7	-0.014	-0.001	0.000	0.000	0.029	0.000	0.000	0.000	0.015	-0.001	0.000	0.000
-0.6	-0.022	-0.015	-0.017	0.000	0.023	-0.012	-0.018	0.000	0.001	-0.027	-0.035	0.000
-0.6	-0.018	-0.010	-0.013	-0.013	0.018	-0.022	-0.030	-0.030	0.000	-0.032	-0.043	-0.042
-0.4	0.002	-0.017	-0.020	-0.023	0.026	0.004	-0.004	0.012	0.028	-0.012	-0.024	-0.011
-0.2	-0.013	0.006	0.000	0.000	0.007	0.006	0.000	0.000	-0.006	0.012	0.000	0.000
-0.2	-0.006	-0.004	-0.007	-0.002	-0.005	-0.001	-0.005	-0.000	-0.011	-0.006	-0.013	-0.002
-0.1	-0.010	0.001	0.000	0.000	-0.012	0.001	0.000	0.000	-0.022	0.002	0.000	0.000
0	-0.016	0.005	0.000	0.000	0.007	0.002	0.000	0.000	-0.009	0.007	0.000	0.000
0	-0.014	0.017	0.012	0.047	-0.005	0.006	0.000	0.021	-0.019	0.023	0.012	0.068
0.1	-0.008	0.010	0.003	0.000	-0.005	0.012	0.007	0.000	-0.013	0.023	0.010	0.000
0.2	-0.015	0.000	0.000	0.000	0.005	0.000	0.000	0.000	-0.010	0.000	0.000	0.000
0.2	0.000	0.043	0.053	-0.003	-0.009	0.028	0.040	-0.014	-0.010	0.072	0.093	-0.016
0.5	-0.012	-0.002	0.000	0.000	-0.036	-0.002	0.000	0.000	-0.048	-0.004	0.000	0.000
0.5	-0.023	0.010	0.016	0.001	-0.041	-0.001	0.004	-0.001	-0.065	0.009	0.021	0.001
0.7	0.011	0.007	0.011	0.000	-0.019	0.002	0.009	0.000	-0.008	0.009	0.020	0.000
0.9	0.006	-0.030	-0.025	0.000	-0.042	-0.031	-0.025	0.000	-0.036	-0.060	-0.050	0.000
0.9	-0.015	-0.004	0.000	0.000	-0.060	-0.006	0.000	0.000	-0.074	-0.010	0.000	0.000
1.2	-0.035	0.047	0.058	0.027	-0.062	0.049	0.073	0.028	-0.097	0.096	0.131	0.056
1.3	-0.009	-0.011	0.000	0.000	-0.068	-0.012	0.000	0.000	-0.077	-0.024	0.000	0.000
1.6	-0.062	0.040	0.049	0.028	-0.134	0.059	0.083	0.038	-0.196	0.099	0.132	0.066
1.7	-0.009	-0.018	0.000	0.000	-0.099	-0.018	0.000	0.000	-0.108	-0.036	0.001	0.000
2.1	-0.013	-0.019	0.001	0.000	-0.121	-0.021	0.000	0.000	-0.134	-0.040	0.001	0.000
2.2	-0.054	-0.017	0.006	-0.023	-0.179	0.009	0.042	0.000	-0.233	-0.009	0.048	-0.023
2.4	-0.005	0.159	0.175	0.130	-0.137	0.155	0.200	0.120	-0.142	0.314	0.375	0.250
2.5	-0.005	0.019	0.001	0.000	-0.140	0.003	0.001	0.000	-0.146	0.022	0.001	0.000
2.6	-0.135	-0.012	0.011	-0.021	-0.258	-0.080	-0.051	-0.073	-0.392	-0.092	-0.040	-0.094
min	**-0.135**	**-0.230**	**-0.233**	**-0.280**	**-0.258**	**-0.227**	**-0.263**	**-0.259**	**-0.392**	**-0.457**	**-0.496**	**-0.539**
max	**0.214**	**0.192**	**0.175**	**0.264**	**0.301**	**0.231**	**0.205**	**0.292**	**0.515**	**0.423**	**0.375**	**0.556**
mean	**-0.009**	**0.004**	**0.003**	**0.000**	**-0.013**	**0.002**	**0.002**	**-0.001**	**-0.022**	**0.006**	**0.005**	**-0.001**
sd	**0.051**	**0.065**	**0.062**	**0.075**	**0.105**	**0.071**	**0.077**	**0.078**	**0.148**	**0.134**	**0.138**	**0.153**

Table 4. RMSE Differences of Item Difficulty When SD are Fixed

	n(0,1)vs.n(0,1) - n(0,1)vs.(0.5,1)				n(0,1)vs.n(0.5,1) - n(0,1)vs.n(1,1)				n(0,1)vs.n(0,1) - n(0,1)vs.(1,1)			
	HLM	BILOG	B-MG	ST	HLM	BILOG	B-MG	ST	HLM	BILOG	B-MG	ST
-2.3	0.622	0.304	-0.009	0.000	0.210	0.326	-0.002	0.000	0.831	0.631	-0.011	0.000
-2.2	0.816	0.539	0.129	0.868	0.221	0.441	0.103	0.702	1.038	0.980	0.231	1.570
-1.9	0.433	0.187	-0.006	0.000	0.097	0.181	-0.001	0.000	0.530	0.368	-0.007	0.000
-1.8	0.408	0.370	0.029	0.650	0.197	0.324	0.041	0.552	0.605	0.694	0.070	1.202
-1.5	0.493	0.169	-0.005	0.000	0.121	0.168	-0.001	0.000	0.614	0.337	-0.006	0.000
-1.4	0.384	0.494	0.048	0.809	0.182	0.439	0.022	0.700	0.565	0.933	0.069	1.510
-1.1	0.297	0.085	-0.002	0.000	0.072	0.074	0.000	0.000	0.369	0.159	-0.003	0.000
-1	0.123	0.294	0.052	0.543	0.055	0.177	-0.031	0.349	0.178	0.472	0.022	0.893
-0.7	-0.160	0.125	0.011	0.000	-0.122	0.091	0.006	0.000	-0.282	0.216	0.018	0.000
-0.7	0.280	0.040	-0.001	0.000	0.067	0.031	0.000	0.000	0.348	0.071	-0.002	0.000
-0.6	0.152	0.092	0.012	0.000	0.045	0.048	-0.004	0.000	0.197	0.141	0.008	0.000
-0.6	0.095	0.133	0.019	0.326	0.077	0.062	0.012	0.192	0.172	0.194	0.031	0.518
-0.4	0.084	0.090	0.008	0.266	0.046	0.021	-0.005	0.130	0.130	0.112	0.003	0.396
-0.2	0.134	0.049	-0.001	0.000	0.035	0.040	0.000	0.000	0.169	0.088	-0.002	0.000
-0.2	0.040	0.051	0.010	0.220	0.018	-0.031	0.003	0.067	0.058	0.020	0.014	0.287
-0.1	-0.063	-0.019	0.000	0.000	-0.032	-0.028	0.000	0.000	-0.095	-0.046	0.001	0.000
0	0.068	-0.050	0.001	0.000	0.018	-0.056	0.000	0.000	0.087	-0.105	0.002	0.000
0	-0.012	0.033	0.017	0.198	-0.008	-0.056	-0.005	0.033	-0.021	-0.023	0.012	0.230
0.1	-0.013	0.048	0.003	0.000	-0.021	0.017	0.003	0.000	-0.034	0.066	0.006	0.000
0.2	0.070	-0.006	0.000	0.000	0.012	-0.017	0.000	0.000	0.082	-0.023	0.000	0.000
0.2	-0.066	-0.171	-0.029	-0.081	-0.067	-0.232	-0.009	-0.212	-0.133	-0.404	-0.038	-0.292
0.5	-0.226	-0.026	0.001	0.000	-0.069	-0.039	0.000	0.000	-0.296	-0.065	0.001	0.000
0.5	-0.056	-0.133	0.002	-0.059	-0.031	-0.195	0.008	-0.172	-0.087	-0.327	0.010	-0.231
0.7	-0.253	-0.161	-0.022	0.000	-0.124	-0.202	-0.023	0.000	-0.377	-0.363	-0.045	0.000
0.9	-0.362	-0.131	0.001	0.000	-0.144	-0.159	0.010	0.000	-0.507	-0.291	0.012	0.000
0.9	-0.391	-0.087	0.002	0.000	-0.109	-0.100	0.000	0.000	-0.499	-0.187	0.003	0.000
1.2	-0.310	-0.328	0.005	-0.333	-0.191	-0.489	-0.070	-0.533	-0.502	-0.816	-0.065	-0.866
1.3	-0.450	-0.100	0.003	0.000	-0.140	-0.119	0.001	0.000	-0.590	-0.219	0.003	0.000
1.6	-0.376	-0.416	0.019	-0.467	-0.187	-0.489	0.029	-0.592	-0.563	-0.906	0.047	-1.058
1.7	-0.671	-0.135	0.004	0.000	-0.184	-0.158	0.001	0.000	-0.855	-0.293	0.004	0.000
2.1	-0.823	-0.189	0.005	0.000	-0.220	-0.212	0.001	0.000	-1.043	-0.401	0.006	0.000
2.2	-0.539	-0.555	-0.028	-0.632	-0.347	-0.607	0.021	-0.746	-0.885	-1.162	-0.007	-1.378
2.4	-0.588	-0.730	-0.088	-0.850	-0.394	-0.851	-0.109	-1.029	-0.982	-1.581	-0.196	-1.879
2.5	-0.923	-0.282	0.007	0.000	-0.261	-0.281	0.001	0.000	-1.183	-0.563	0.009	0.000
2.6	-0.533	-0.709	-0.085	-0.819	-0.279	-0.749	-0.007	-0.926	-0.812	-1.458	-0.092	-1.745
min	**-0.923**	**-0.730**	**-0.088**	**-0.850**	**-0.394**	**-0.851**	**-0.109**	**-1.029**	**-1.183**	**-1.581**	**-0.196**	**-1.879**
max	**0.816**	**0.539**	**0.129**	**0.868**	**0.221**	**0.441**	**0.103**	**0.702**	**1.038**	**0.980**	**0.231**	**1.570**
mean	**-0.066**	**-0.032**	**0.003**	**0.018**	**-0.042**	**-0.075**	**0.000**	**-0.042**	**-0.108**	**-0.107**	**0.003**	**-0.024**
sd	**0.408**	**0.288**	**0.035**	**0.364**	**0.155**	**0.296**	**0.031**	**0.367**	**0.556**	**0.583**	**0.060**	**0.728**

Table 5. Mean and SD of Item Difficulty RMSEs

	n(0,1)				n(0,.75)				n(0,0.5)			
	HLM	**BILOG**	**B-MG**	**ST**	**HLM**	**BILOG**	**B-MG**	**ST**	**HLM**	**BILOG**	**B-MG**	**ST**
min	0.021	0.006	0.006	0.007	0.031	0.006	0.006	0.007	0.043	0.006	0.006	0.007
max	3.645	4.307	4.301	4.258	3.658	4.247	4.302	4.258	3.543	4.197	4.302	4.258
mean	1.013	0.962	0.973	0.978	1.022	0.958	0.969	0.977	1.036	0.956	0.967	0.978
sd	1.116	1.189	1.217	1.152	1.109	1.178	1.209	1.156	1.114	1.174	1.203	1.165
	n(0.5,1)				**n(0.5,0.75)**				**n(0.5, 0.5)**			
	HLM	**BILOG**	**B-MG**	**ST**	**HLM**	**BILOG**	**B-MG**	**ST**	**HLM**	**BILOG**	**B-MG**	**ST**
min	0.043	0.007	0.006	0.007	0.043	0.010	0.006	0.007	0.039	0.009	0.006	0.007
max	4.255	4.219	4.310	4.258	4.235	4.221	4.311	4.258	4.294	4.221	4.312	4.258
mean	1.080	0.994	0.969	0.959	1.076	0.990	0.965	0.957	1.083	0.987	0.963	0.957
sd	1.206	1.285	1.227	1.169	1.211	1.268	1.211	1.164	1.235	1.258	1.204	1.168
	n(1,1)				**n(1,0.75)**				**n(1,0.5)**			
	HLM	**BILOG**	**B-MG**	**ST**	**HLM**	**BILOG**	**B-MG**	**ST**	**HLM**	**BILOG**	**B-MG**	**ST**
min	0.032	0.003	0.006	0.007	0.034	0.004	0.006	0.007	0.038	0.008	0.006	0.007
max	4.516	4.928	4.312	4.258	4.537	4.635	4.314	4.258	4.500	4.535	4.316	4.258
mean	1.130	1.071	0.969	1.004	1.133	1.067	0.965	1.002	1.045	0.956	0.890	0.909
sd	1.290	1.438	1.231	1.303	1.304	1.419	1.213	1.296	1.225	1.274	1.143	1.197

Table 6: Mean and SD of Person RMSEs

	n(0,1)				n(0,0.75)				n(0,0.5)			
	HLM	**BILOG**	**B-MG**	**ST**	**HLM**	**ILOG**	**B-MG**	**ST**	**HLM**	**ILOG**	**B-MG**	**ST**
min	0.0000	0.0264	0.0018	0.0114	0.0000	0.0311	0.0018	0.0117	0.0000	0.0314	0.0017	0.0104
max	8.6562	7.6347	5.9493	7.6204	7.3692	7.5289	5.9493	7.6204	7.3692	7.4368	5.9493	7.6204
mean	1.0572	1.0086	0.6980	0.7652	0.9330	0.8922	0.5874	0.6546	0.8406	0.8064	0.5079	0.5755
sd	1.3948	1.2465	0.9945	1.0410	1.2288	1.0848	0.8741	0.9908	1.1381	0.9918	0.8224	0.9903
	n(0.5,1)				**n(0.5,0.75)**				**n(0.5,0.5)**			
	HLM	**BILOG**	**B-MG**	**ST**	**HLM**	**ILOG**	**B-MG**	**ST**	**HLM**	**ILOG**	**B-MG**	**ST**
min	0.0000	0.0311	0.0024	0.0116	0.0001	0.0283	0.0012	0.0122	0.0000	0.0244	0.0016	0.0115
max	10.8616	8.0418	5.9493	7.6204	7.5417	7.1395	5.9493	7.6204	7.3692	7.0385	5.9493	7.6204
mean	1.1361	1.0603	0.6484	0.8639	1.0095	0.9419	0.5369	0.7505	0.9145	0.8523	0.4564	0.6689
sd	1.5009	1.3123	0.9354	1.1171	1.3178	1.1472	0.8423	1.0167	1.2111	1.0512	0.8156	0.9783
	n(1,1)				**n(1,0.75)**				**n(1,0.5)**			
	HLM	**BILOG**	**B-MG**	**ST**	**HLM**	**ILOG**	**B-MG**	**ST**	**HLM**	**ILOG**	**B-MG**	**ST**
min	0.0001	0.0225	0.0018	0.0124	0.0000	0.0206	0.0023	0.0136	0.0001	0.0170	0.0024	0.0148
max	13.3170	9.8622	6.2197	8.7569	9.6085	7.3191	5.9493	7.6204	7.3692	6.5917	5.9493	7.6204
mean	1.3407	1.1857	0.6853	1.0878	1.2116	1.0606	0.5709	0.9722	1.1141	0.9664	0.4875	0.8870
sd	1.7781	1.4969	0.9705	1.2997	1.5518	1.3112	0.8586	1.1059	1.4018	1.1945	0.8177	0.9844

Figure 1: Item RMSE: n(0,1) vs. n(0,1)

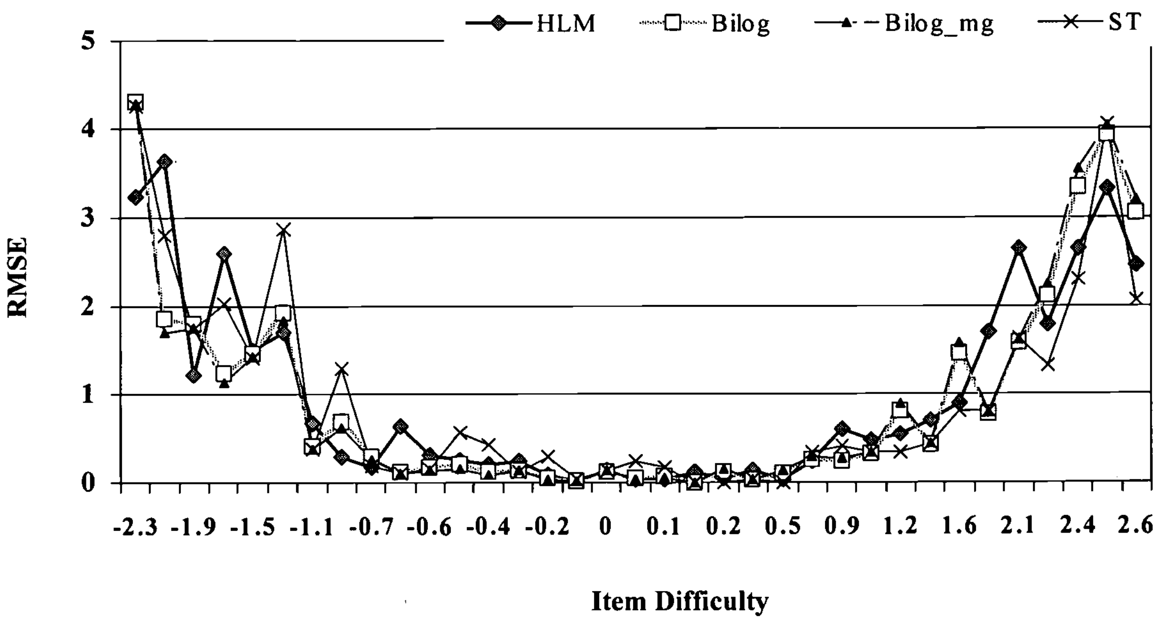


Figure 2: Item RMSE: n(0,1) vs. n(0,.75)

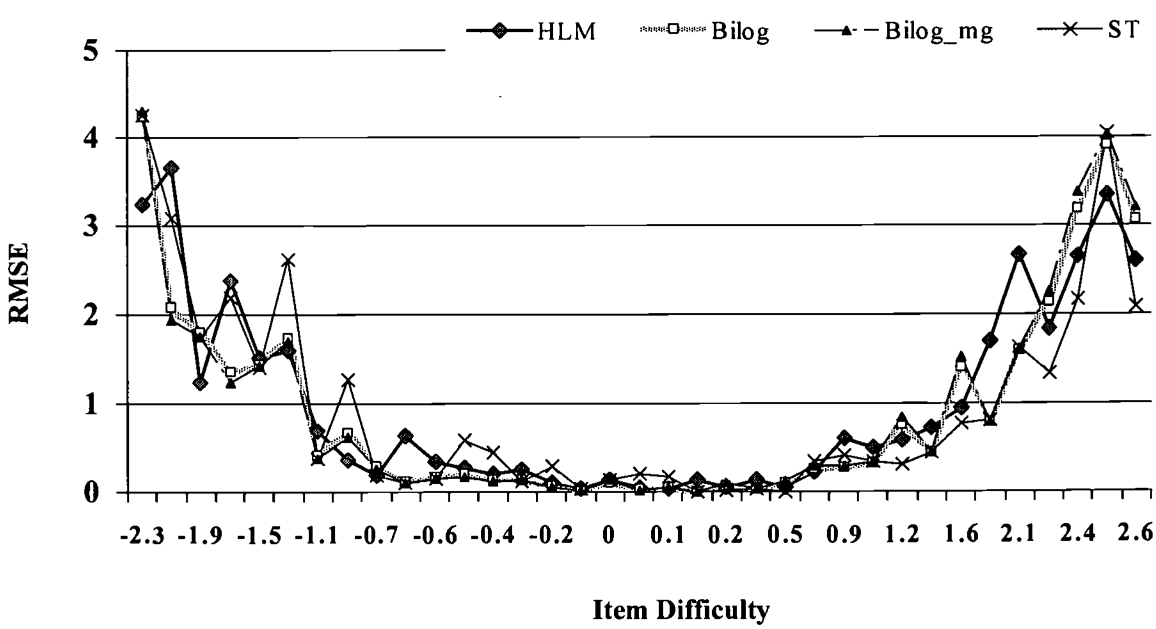


Figure 3: Item RMSE: n(0,1) vs. n(0,.5)

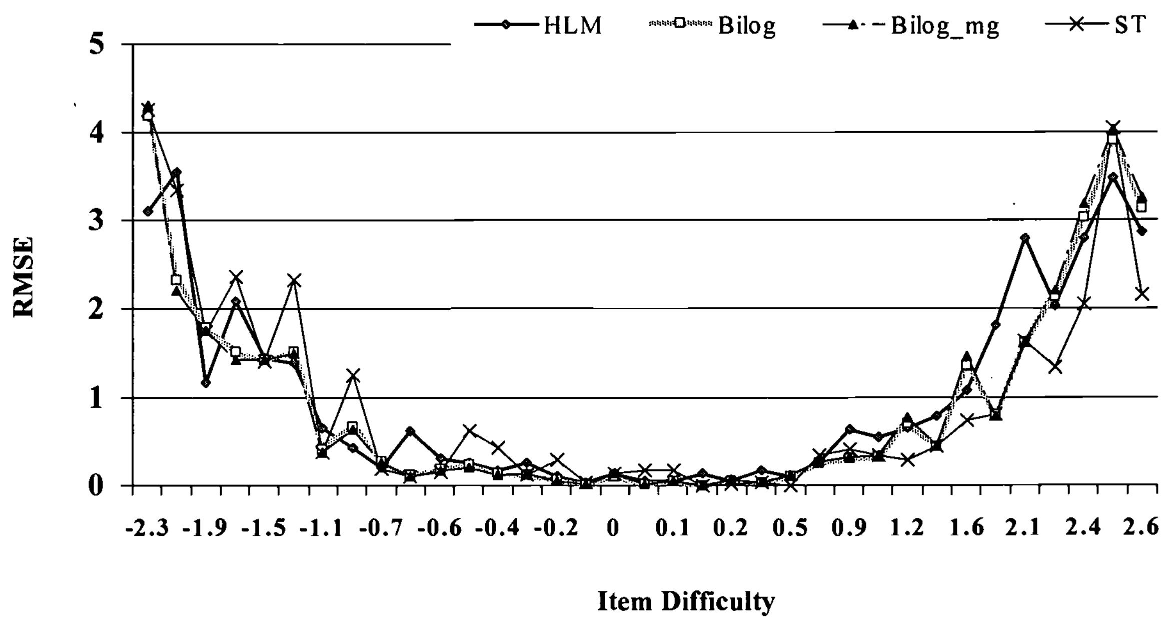


Figure 4: Item RMSE: n(0,1) vs. n(.5,1)

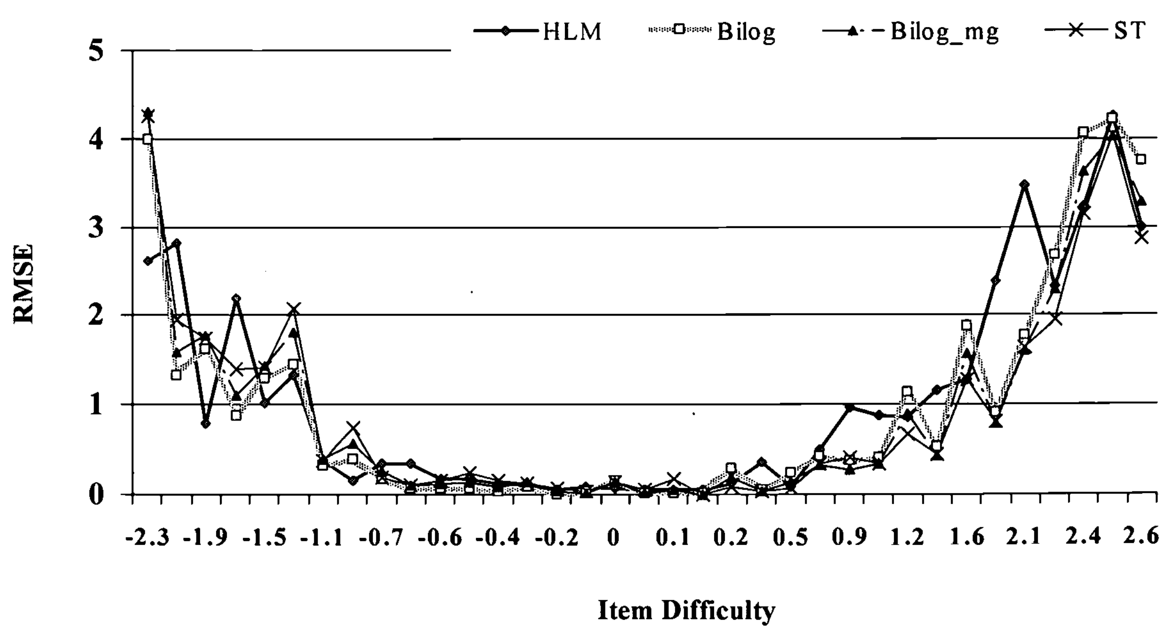


Figure 5: Item RMSE: n(0,1) vs. n(.5,.75)

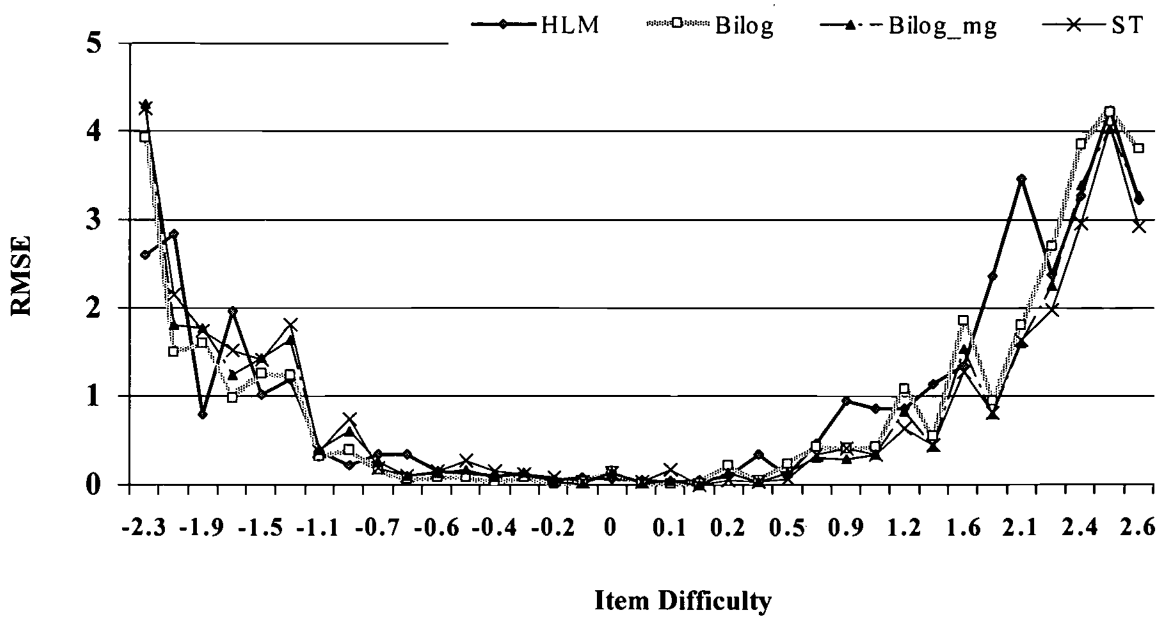


Figure 6: Item RMSE: n(0,1) vs. n(.5,.5)

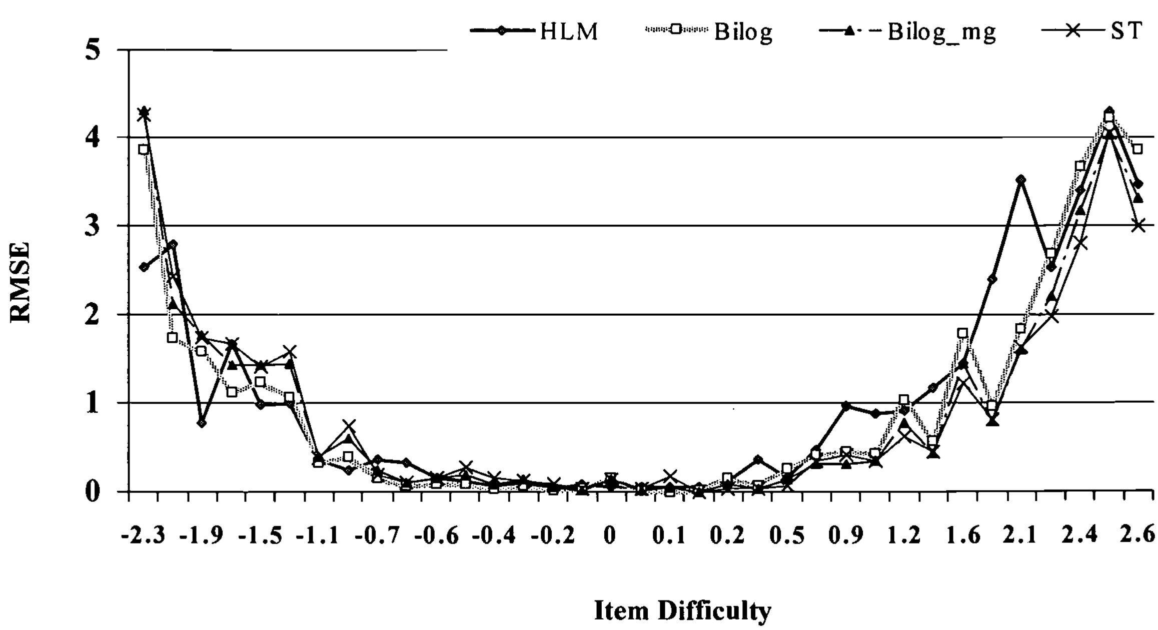


Figure 7: Item RMSE: n(0,1) vs. n(1,1)

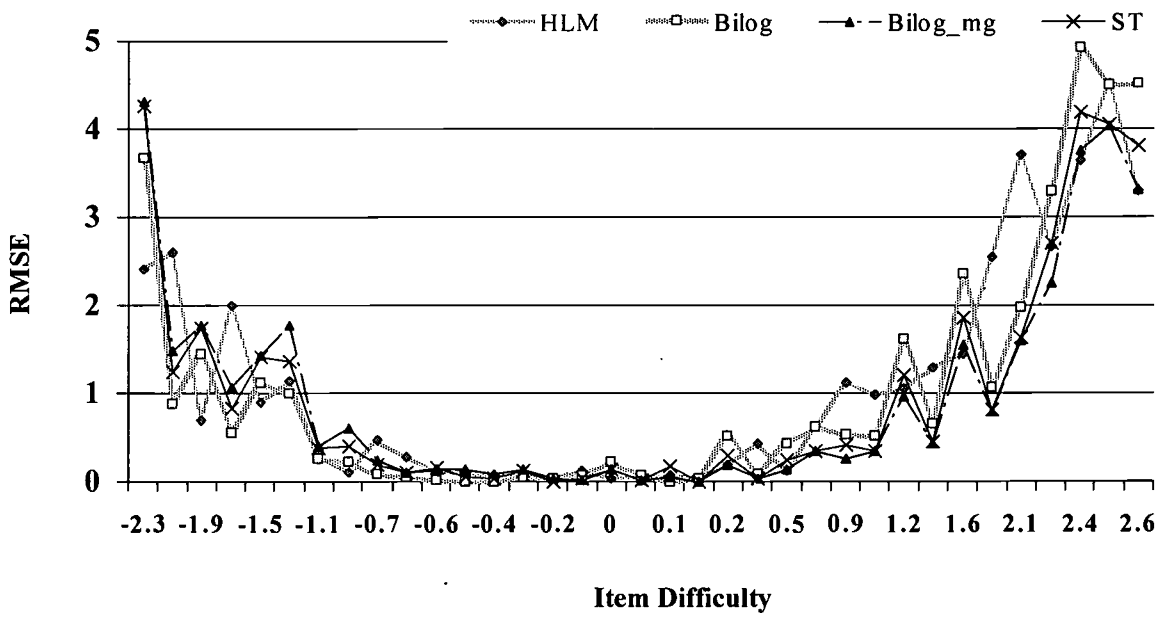


Figure 8: Item RMSE: n(0,1) vs. n(1,.75)

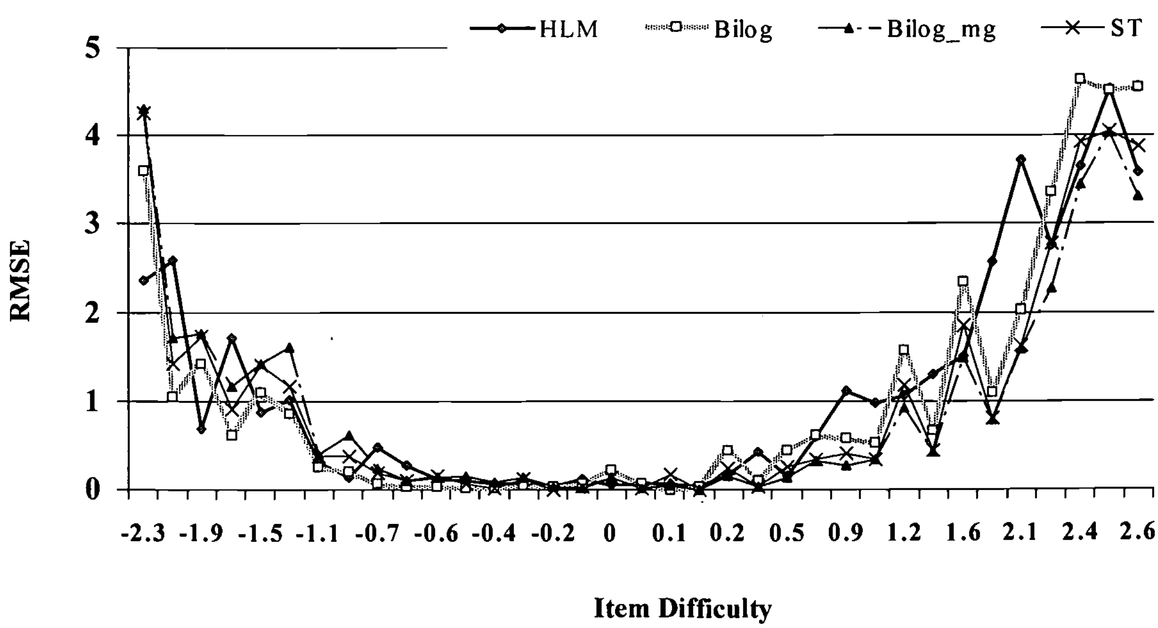


Figure 9: Item RMSE: n(0,1) vs. n(1,.5)

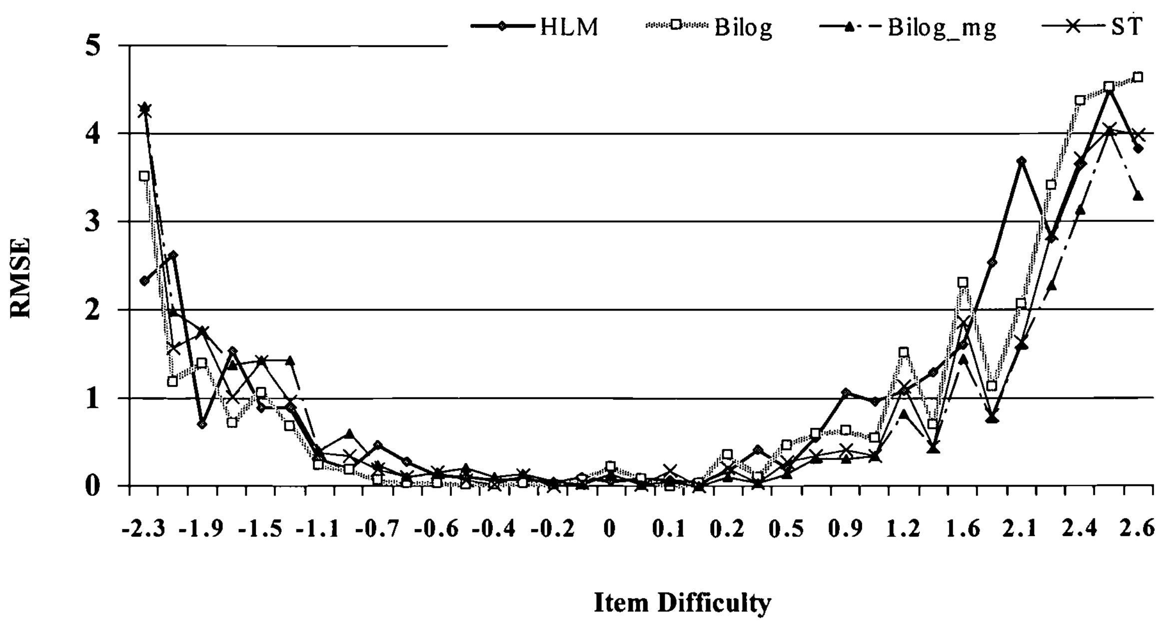


Figure 10: Person RMSEs; N(0,1) vs. N(0,1)

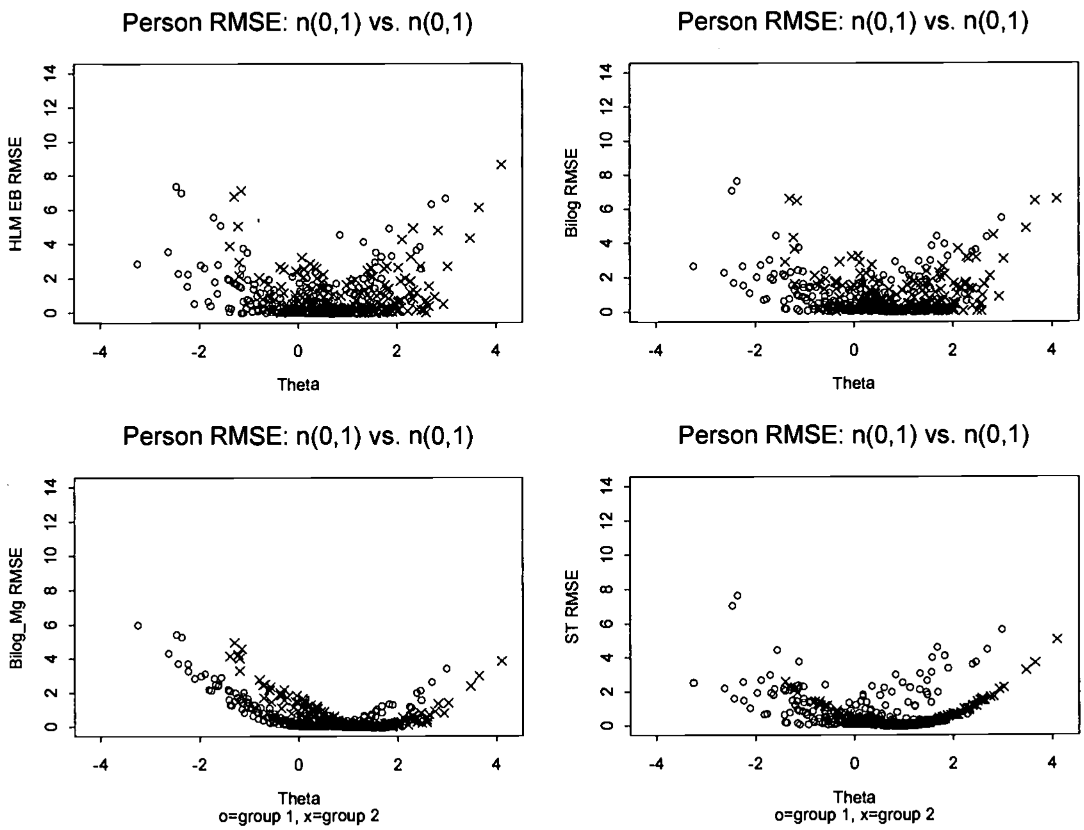


Figure 11: Person RMSEs; N(0,1) vs. N(0,0.75)

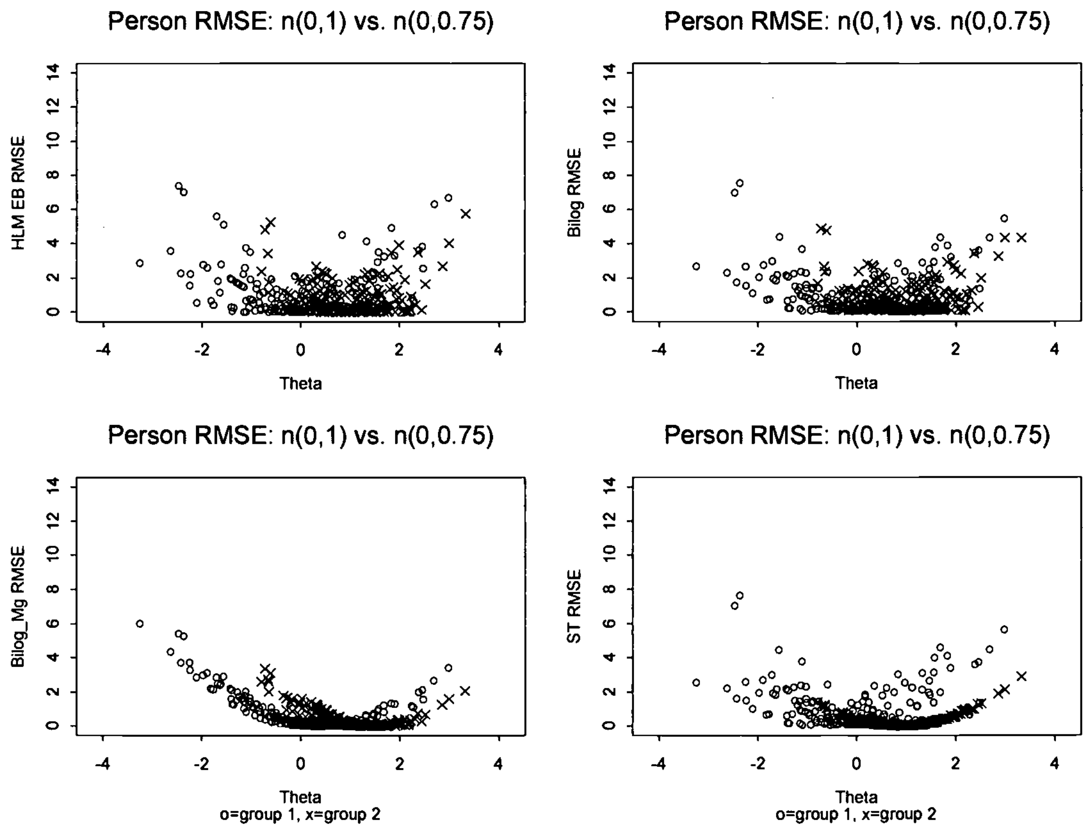
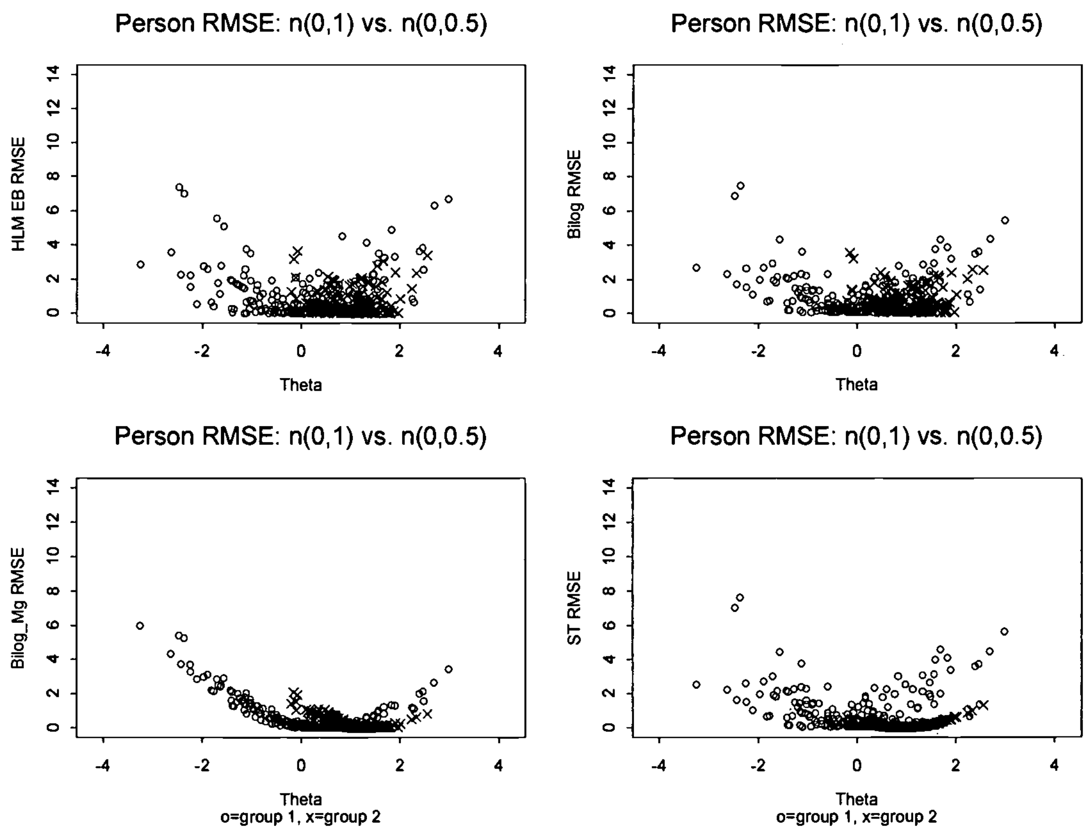
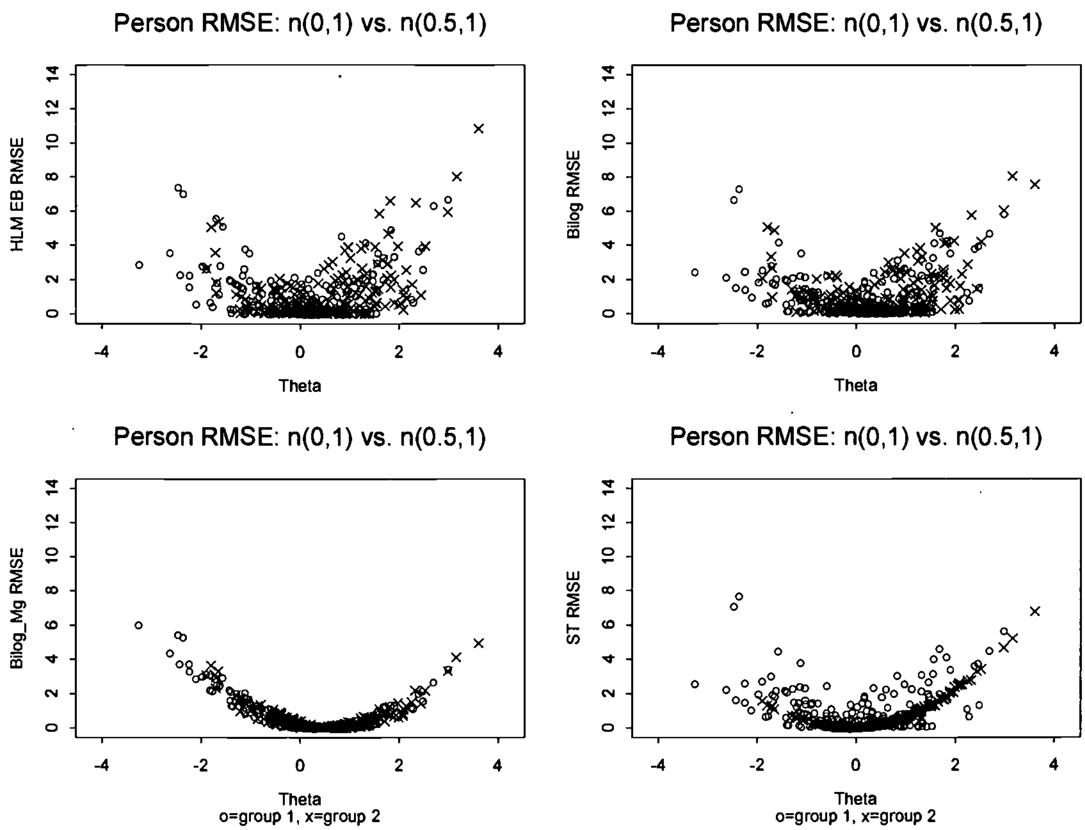


Figure 12: Person RMSEs; N(0,1) vs. N(0,0.5)



29

Figure 13: Person RMSEs; N(0,1) vs. N(0.5, 1)



30

Figure 14: Person RMSEs; N(0,1) vs. N(0.5, 0.75)

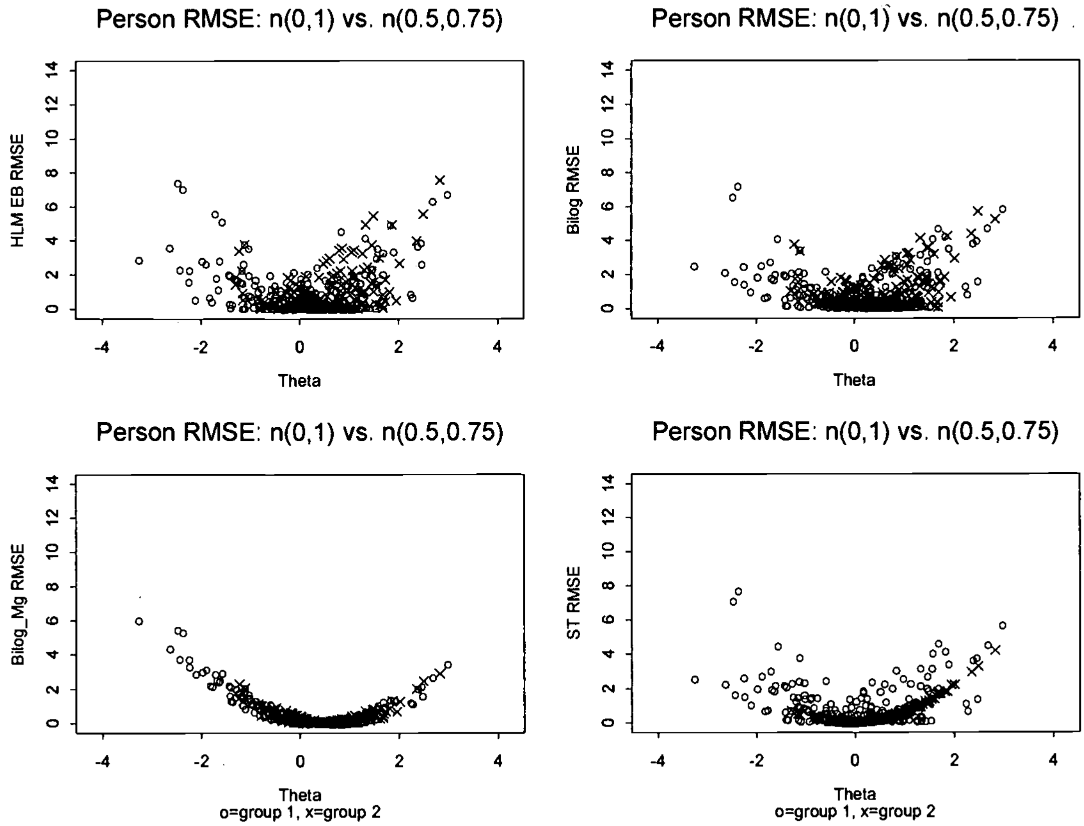


Figure 15: Person RMSEs; N(0,1) vs. N(0.5, 0.5)

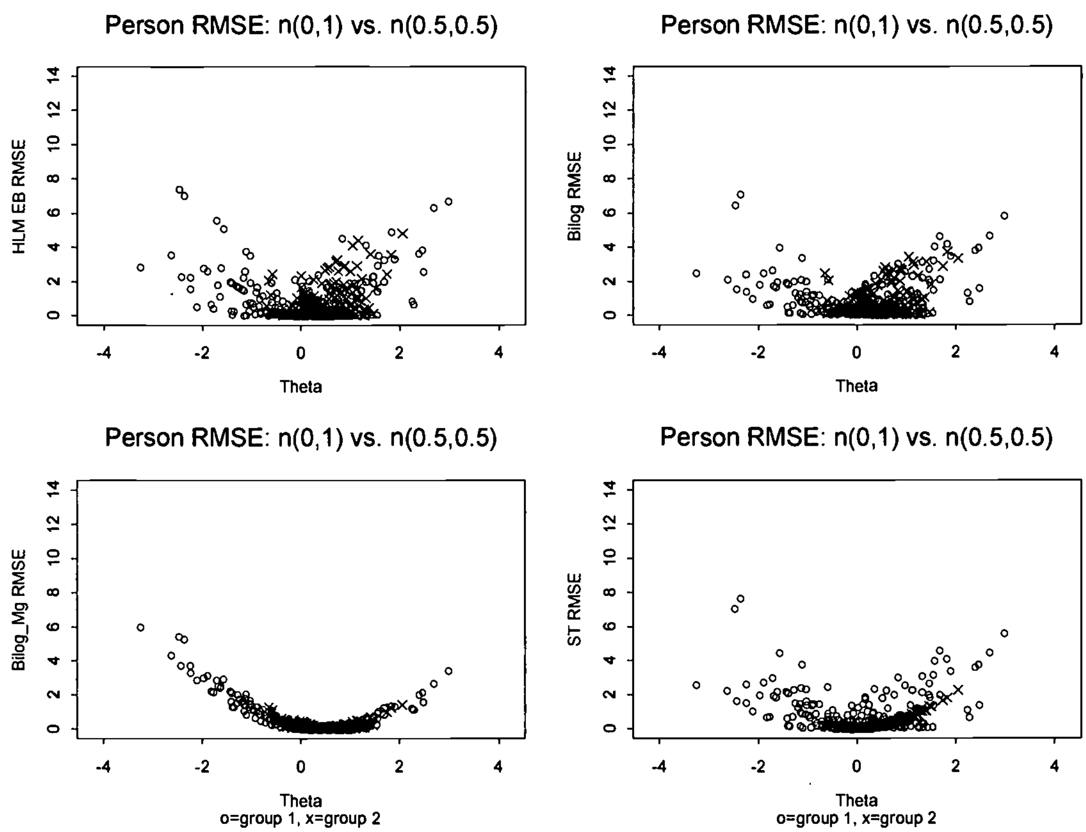


Figure 16: Person RMSEs; N(0,1) vs. N(1, 1)

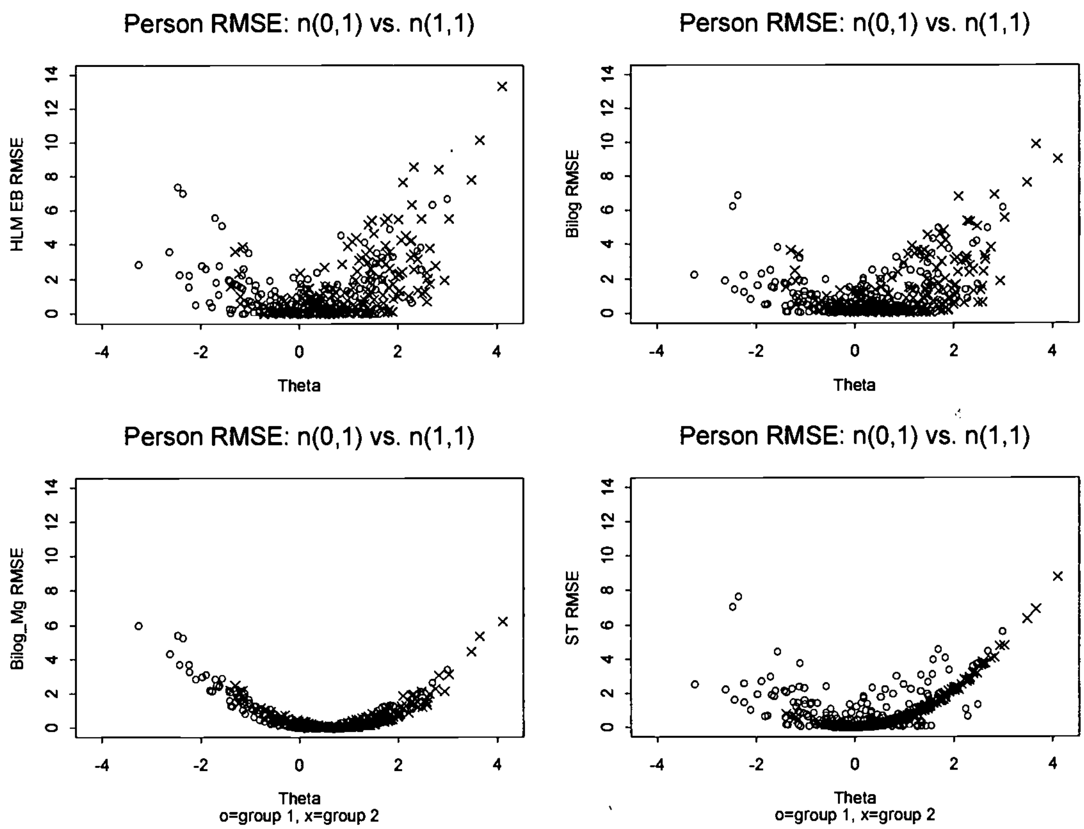


Figure 17: Person RMSEs; N(0,1) vs. N(1, 0.75)

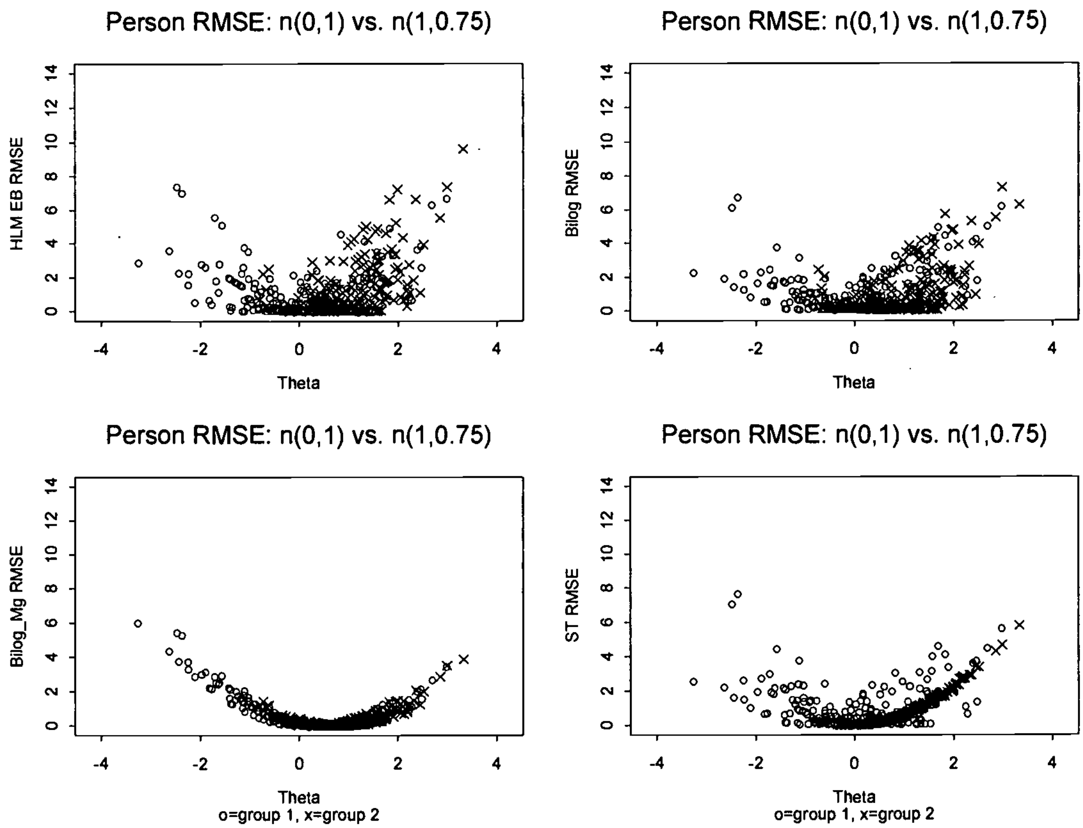
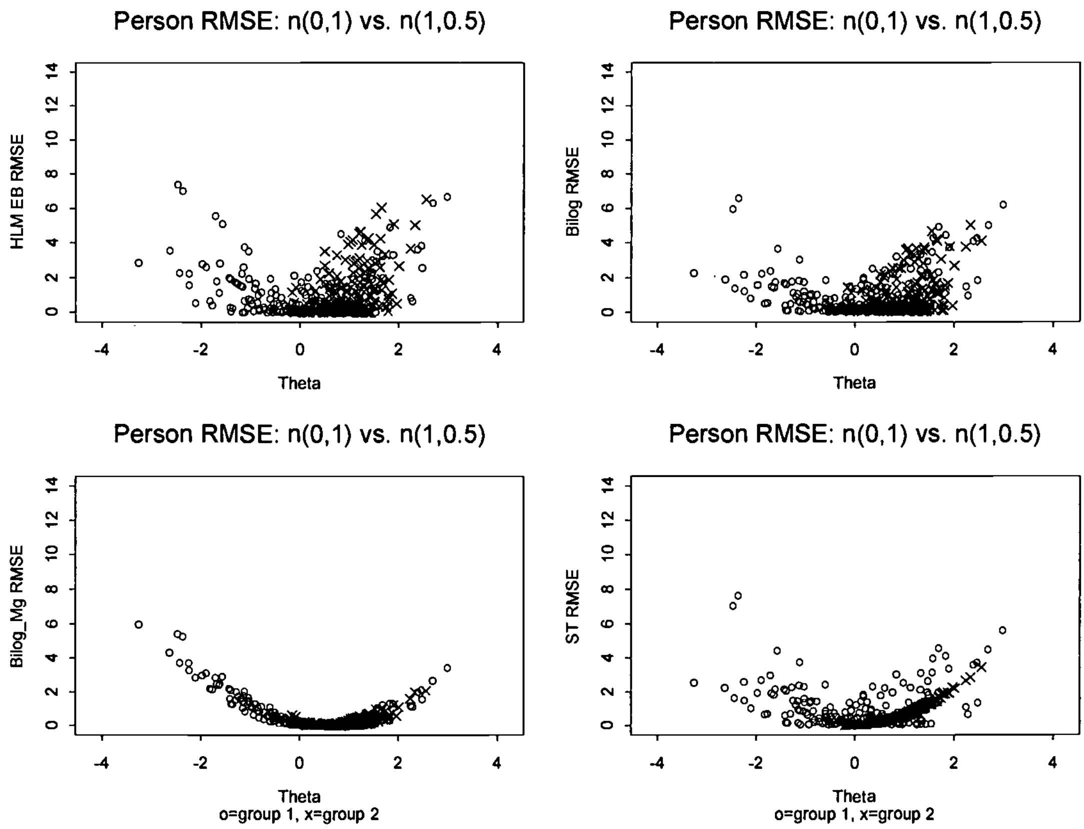


Figure 18: Person RMSEs; N(0,1) vs. N(1, 0.5)



ERIC® **Clearinghouse on Assessment and Evaluation**

University of Maryland
1129 Shriver Laboratory
College Park, MD 20742-5701

Tel: (800) 464-3742
(301) 405-7449
FAX: (301) 405-8134
ericae@ericae.net
http://ericae.net

May 8, 2000

Dear AERA Presenter,

Hopefully, the convention was a productive and rewarding event. As stated in the AERA program, presenters have a responsibility to make their papers readily available. If you haven't done so already, please submit copies of your papers for consideration for inclusion in the ERIC database. We are interested in papers from this year's AERA conference and last year's conference. If you have submitted your paper, you can track its progress at http://ericae.net.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the **2000 and 1999 AERA Conference**. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

Please sign the Reproduction Release Form enclosed with this letter and send **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can mail your paper to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

AERA 2000/ERIC Acquisitions
University of Maryland
1129 Shriver Laboratory
College Park, MD 20742

Sincerely,

Lawrence M. Rudner

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

ERIC is a project of the Department of Measurement, Statistics & Evaluation