A Comparison of Scoring Options for Omitted and Not-Reached Items Through the Recovery of IRT Parameters When Utilizing the Rasch Model and Joint Maximum Likelihood Estimation

Michael Custer, Sid Sharairi, and David Swift

Riverside Publishing Company

Paper Presented at the Annual Meeting of the

National Council on Measurement in Education

Vancouver, British Columbia

April 12-16

2012

Please send correspondence regarding this paper to Michael.Custer@hmhpub.com

Abstract

This paper utilized the Rasch model and Joint Maximum Likelihood Estimation to study different scoring options for omitted and not-reached items. Three scoring treatments were studied. The first method treated omitted and not-reached items as "ignorable/blank". The second treatment, scored omits as incorrect with "0" and left not-reached as blank and the third treatment scored both omitted and not-reached items as incorrect with "0". These scoring treatments were studied across two levels of missing data: ".81% omit rate/10% not-reached" and "1.62% omit rate/20% not-reached" and two levels of not reaching the end of the test: independent of ability and dependent on ability. In terms of Bias, RMSD, and the number of outlier differences between estimated and "true" parameters, performance was best when omitted and not-reached items were ignored or left blank. Performance was poorest when omitted and not-reached items were scored as incorrect with "0".

A Comparison of Scoring Options for Omitted and Not-Reached Items Through the Recovery of IRT Parameters When Utilizing the Rasch Model and Joint Maximum Likelihood Estimation

Background

Several decisions are required before test items and examinee abilities can be calibrated. Choices regarding the most appropriate Item Response Theory (IRT) model, the IRT software and estimation method to be utilized and the most appropriate scoring treatment for omitted and not-reached items are some of the most important decisions to be made. The primary focus of this paper is the study and comparison of scoring options for omitted and not-reached items when the Rasch model and Joint Maximum Likelihood Estimation (JMLE) are utilized.

By definition, an omitted item is one which an examinee chooses not to answer because he or she feels that they lack sufficient knowledge to answer the item correctly. Instead of guessing at a response, the examinee chooses to leave the item unanswered. In contrast not-reached items are those that an examinee does not respond to because they are unable to finish the test in the allotted time.

Through the utilization of models incorporating IRT, investigators have studied the optimal method for scoring missing data not only in terms of the scoring option chosen but also in terms of the estimation method used. Joint Maximum Likelihood Estimation (JMLE) and Marginal Maximum Likelihood Estimation (MMLE) are two common estimation methods used within IRT. With JMLE, item and ability parameters are

estimated simultaneously with an item's difficulty estimate being dependent upon only those examinees who responded to that particular item and an examinee's ability estimate being dependent upon only the set of items that were responded to by that examinee. In contrast, MMLE treats the ability distribution as known and uses this distribution in the estimation of the item parameters. Once the item parameters are estimated, the examinee abilities are then directly estimated.

Several studies (Lord 1974, Mislevy & Wu 1988, DeAyala, Plake & Impara 2001) have investigated the use of MMLE in the treatment of missing data and have recommended that omitted items should not be ignored or left missing in their scoring treatment. These studies also note that scoring omitted items as incorrect (0) is problematic since even the lowest ability examinee has some probability of getting an omitted item correct. The general recommendation is to score omitted items as fractionally correct. With respect to not-reached items, Lord (1980) and Mislevy (1988) suggested that these items as opposed to omitted items could be ignored in their scoring treatment as long the items are answered in sequential order and the test is "nearly non-speeded".

In studies utilizing the Rasch Model and JMLE, Ludlow & Leary (1999) found that treating omitted items as "ignorable/blank" inflated the directly estimated abilities for examinees who omitted items and deflated the item difficulties for items with omitted data. Likewise, scoring not-reached items as incorrect with a 0, deflated the directly estimated abilities for examinees who did not reach the end of the test and inflated the item difficulties for items positioned at the end of the test. A two stage process was recommended that first incorporated an item calibration with omitted items being treated

as incorrect and not-reached items scored as blank/missing. This would be followed by a person calibration with omitted and not-reached items scored as incorrect with a zero. Shin (2009) studied the impact of different scoring methods on IRT-based true score equating. The study recommended that omitted and not-reached items should be ignored and left blank and should not be scored as incorrect. The study also found that the benefits of treating omitted and not-reached items as "ignorable/blank" increased as the sample size increased.

Demars (2003) utilized the one-parameter model (1PL) and both JMLE and MMLE to investigate the potential link between ability differences and not-reached rates and found that when not-reached items were left blank/missing and not-reached rates were independent of ability, both JMLE and MMLE produced unbiased item parameter estimates. However, when not-reached rates were linked to examinee ability, JMLE produced unbiased item parameter estimates while MMLE produced difficulty estimates that were consistently lower than the true item parameter estimates. DeMars concluded that although JMLE theory allows for ignoring missing data, MMLE does so only if not-reached rates are random or independent of examinee ability.

In contrast to the work described above, Koretz, Lewis, Skewes-Cox, and Burstein (1993) utilized data from the 1990 NAEP mathematics assessment to study the relationship between missing data and item as well as population group characteristics. They found that differences in omit and not-reached rates were partially attributable to ability/proficiency differences.

This study implements the Rasch model as operationalized through WINSTEPS 3.65 which utilizes JMLE (also called UCON). With the Rasch Model the likelihood

function is conditioned on the raw score as a sufficient statistic for the ability parameter. This is reflected in a subtle difference in the manner in which abilities are estimated relative to IRT models in general. With the Rasch model, ability estimates are adjusted until the difference between an examinee's expected score and his or her observed raw score is sufficiently small. With IRT models, ability estimates are continually adjusted until the difference between an examinee's computed item probabilities and his or her item response vector is negligibly small. In both cases, ability estimation is discontinued once a logit convergence criterion has been met.

Three different scoring methods were investigated. First, omitted and not-reached items were ignored or left blank, second, they were scored as incorrect with item scores of zero and third, omitted items were scored as incorrect and not-reached items were left blank. These three scoring methods were evaluated under two conditions. The first being when not-reached rates were independent of ability. In this case whether an examinee reached the end of the test or not was simply a matter of random selection. The second condition was when not-reached rates were dependent on ability with lower ability examinees having a greater likelihood of not reaching the end of the test. Lastly, these three scoring methods with not-reached being independent and dependent on ability were evaluated across two levels of missing data (".81% omit rate\10% not-reached" and "1.62% omit rate\20% not-reached").

Method

The basic simulation and calibration process involved the creation and calibration of 24 data sets. This simulation and calibration process which is described below was executed 10 times for a total of 240 data sets calibrated across 240 runs.

Data were first simulated for 500 examinees across 40 items using WINGEN2. Two data sets were simulated. The first data set was simulated utilizing the 1PL with the mean and standard deviation (SD) for items and abilities set to 0 and 1 respectively and the second data set utilized the two-parameter model (2PL) with the mean and SD for items and abilities set to 0 and 1 accompanied by an item discrimination parameter modeled to have a mean of 1 with an SD of .25. In both cases the items were sequenced from easiest to hardest before response data were simulated. These "complete" data sets were then calibrated with the resulting item and ability parameter estimates being treated as the "true" parameter estimates.

Omitted and not-reached data were then programmatically embedded to create additional data sets to include two levels of "missing" data. The first level (".81% omit rate\10% not-reached") incorporated an omit rate of .81% that was modeled after a paper-pencil format version of a Math and Reading/ELA test that was administered by a large school district in the southwest United States at grades 3-12 during the fall of 2010. The 10% not-reached rate was modeled after a paper-pencil format version of the Math and Reading components of a commonly used standardized test that is administered at grades 3-8. For the (".81% omit rate\10% not-reached") level, each item was omitted an average of 4 times (0.81% of all examinees) with 10% of the examinees not reaching the end of test. For the second level both the omit and not-reached rates were doubled so that a more extreme "missing" condition could be studied (1.62% omit rate\20% not-reached"). In this second level, each item was omitted an average of 8 times (1.62% of all examinees) with 20% of the examinees not reaching the end of the test (Appendix tables A.1 and A.3).

Items were selected for omission randomly and were programmatically embedded according to the notion that omits could also be guessed at. Hence item data that had originally been scored as 1 was selected for omission 25% of the time and item data that had been originally scored as 0 was selected for omission 75% of the time.

Not-reached items were embedded in two ways (Appendix tables A.2, A.4 and A.5). The first method utilized a random selection of examinees to sequentially embed not-reached item strings between items 33 and 40 by programmatically assigning blanks to item response strings extending from a given item within the 33-40 range to the end of the test. Not-reached items embedded in this manner permitted the study of item and ability estimation when not-reaching the end of the test was independent of ability (random selection). The second method utilized a systematic selection of examinees to assign not-reached item strings so that examinees with a lower ability had a greater probability of not reaching the end of the test. Not-reached items embedded in this manner permitted the study of item and ability estimation when not-reaching the end of the test was dependent on ability (Systematic see A.5). As a result, 8 data sets were created: 2 model methods used to simulate the data x 2 levels of "missing" (".81% omit/10% not-reached" or "1.62% omit/20% not-reached") x 2 methods for determining not-reached (random or systematic).

Each of the 8 data sets were then programmatically scored with the three different scoring options for the omitted and not-reached data. Once the omitted and not-reached items were scored, the 24 data files (8 x 3 scoring options for missing data) were then calibrated with the Rasch variant of the one-parameter model (1PL) using WINSTEPS.

The above described data simulation and calibration process was executed 10 times. Results were analyzed utilizing descriptive statistics, correlational analyses between estimated and true parameters, as well as an averaging, across the 10 runs, of statistics such as the Root Mean Squared Deviation (RMSD) and Bias to measure the precision/quality of the parameter estimation relative to the true item and ability parameters. An average across the 10 runs of the absolute differences between estimated and true ability parameter estimates is also reported.

Results

Descriptive statistics for the initial run are presented in Table 1. These results are only presented for the first run because the descriptive statistics simply did not change by any meaningful amount across the 9 subsequent runs. Relative to the true parameters, the 24 calibrations yielded similar means and standard deviations for the estimated item and ability parameters. The Pearson correlations between the estimated and true ability parameters ranged between .995 and .999. Though not reported, the correlations between the estimated and true item parameters were 1.00.

Table 1) Descriptive Statistics & Correlations For the Initial/First Run

.81% Omit Rate - 10% Not Reached	True Item Mean	True Item SD	Est. Item Mean	Est. Item SD	True Ability Mean	True Ability SD	Est. Ability Mean	Est. Ability SD	Ability Pearson Corr.
1PL Simul. Random NR, Score=Blank	0.000	1.045	0.000	1.048	0.008	1.134	0.020	1.141	.999
1PL Simul. Random NR, Score = (Omit=Zero:NR=Blank)	0.000	1.045	0.000	1.045	0.008	1.134	0.001	1.134	.999
1PL Simul. Random NR, Score=Zero	0.000	1.045	0.000	1.069	0.008	1.134	-0.015	1.128	.998
1PL Simul. Systematic NR,Score=Blank	0.000	1.045	0.000	1.050	0.008	1.134	0.018	1.138	.999
1PL Simul. Systematic NR, Score = (Omit=Zero:NR=Blank)	0.000	1.045	0.000	1.048	0.008	1.134	-0.001	1.132	.999
1PL Simul. Systematic NR, Score=Zero	0.000	1.045	0.000	1.060	0.008	1.134	-0.007	1.137	.999
2PL Simul. Random NR, Score=Blank	0.000	1.021	0.000	1.025	0.037	1.066	0.043	1.068	.999
2PL Simul. Random NR, Score = (Omit=Zero:NR=Blank)	0.000	1.021	0.000	1.025	0.037	1.066	0.025	1.064	.998
2PL Simul. Random NR, Score=Zero	0.000	1.021	0.000	1.045	0.037	1.066	0.011	1.065	.997
2PL Simul. Systematic NR,Score=Blank	0.000	1.021	0.000	1.022	0.037	1.066	0.044	1.068	.999
2PL Simul. Systematic NR, Score = (Omit=Zero:NR=Blank)	0.000	1.021	0.000	1.021	0.037	1.066	0.027	1.063	.999
2PL Simul. Systematic NR, Score=Zero	0.000	1.021	0.000	1.034	0.037	1.066	0.019	1.069	.999

1.62% Omit Rate - 20% Not Reached	True Item Mean	True Item SD	Est. Item Mean	Est. Item SD	True Ability Mean	True Ability SD	Est. Ability Mean	Est. Ability SD	Ability Pearson Corr.
1PL Simul. Random NR, Score=Blank	0.000	1.045	0.000	1.055	0.008	1.134	0.023	1.141	.997
1PL Simul. Random NR, Score = (Omit=Zero:NR=Blank)	0.000	1.045	0.000	1.052	0.008	1.134	-0.013	1.133	.997
1PL Simul. Random NR, Score=Zero	0.000	1.045	0.000	1.096	0.008	1.134	-0.045	1.120	.995
1PL Simul. Systematic NR,Score=Blank	0.000	1.045	0.000	1.049	0.008	1.134	0.022	1.130	.998
1PL Simul. Systematic NR, Score = (Omit=Zero:NR=Blank)	0.000	1.045	0.000	1.046	0.008	1.134	-0.015	1.121	.998
1PL Simul. Systematic NR, Score=Zero	0.000	1.045	0.000	1.076	0.008	1.134	-0.030	1.129	.997
2PL Simul. Random NR, Score=Blank	0.000	1.021	0.000	1.022	0.037	1.066	0.054	1.066	.998
2PL Simul. Random NR, Score = (Omit=Zero:NR=Blank)	0.000	1.021	0.000	1.019	0.037	1.066	0.018	1.051	.997
2PL Simul. Random NR, Score=Zero	0.000	1.021	0.000	1.057	0.037	1.066	-0.007	1.056	.996
2PL Simul. Systematic NR,Score=Blank	0.000	1.021	0.000	1.030	0.037	1.066	0.050	1.073	.998
2PL Simul. Systematic NR, Score= (Omit=Zero:NR=Blank)	0.000	1.021	0.000	1.027	0.037	1.066	0.014	1.059	.998
2PL Simul. Systematic NR, Score=Zero	0.000	1.021	0.000	1.051	0.037	1.066	0.000	1.067	.997

Tables 2 and 3 present an average across all 10 runs with respect to the item RMSD, ability Bias, ability RMSD, and the number of examinee ability measures that differ from their true measure by a value of 0.30 or more. Table 2 presents this information for the ".81% omit rate/10% not-reached" condition and Table 3 for the "1.62% omit rate/20% not-reached" condition. The RMSD is computed by summing the squared differences between estimated and true measures and dividing this value by the number of items (item RMSD) or examinees (ability RMSD) and taking the square root. The Bias statistic is derived by obtaining the difference between estimated and true measures, summing these differences and dividing by the number of examinees. Typically taking an average of the Bias statistic would not be recommended because of sign changes. However, there were no sign changes across the 10 runs hence an averaging became plausible. Given that examinee abilities were simulated to have a mean of 0 and standard deviation of 1, differences between estimated and true abilities that were greater than or equal to 0.30 were considered large enough to indicate a significant drift away from the true ability and are reported in the last column.

As presented in tables 2 and 3, the RMSD for items was lowest and very similar for the scoring treatments where omits and not reached were "ignored/left blank" or when "omits were scored with a 0 and not-reached left blank". The zero fill score option (omit and not-reached="0") yielded item RMSDs that were 2-3 times higher than the other two treatments. This held across both the ".81% omit rate/10% not-reached" and the "1.62% omit rate/20% not-reached" conditions. Under both of these conditions, Bias was smallest when omitted and not-reached items were "ignored/left blank".

Table 2) Summary of Average RMSD, Bias, and Outlier Abilities Across 10 Data Simulations / WINSTEPS Executions for the "81% Omit Rate / 10% Not-Reached Condition"

	RMSD Items	Ability Bias	RMSD Abilities	# of Absolute Differences between Estimated and True Abilities >= 0.30
1PL Simulation - Random NR, Score Omit/NR=blank	0.015	0.008	0.059	3.20
1PL Simulation - Random NR, Score Omit=zero, NR=blank	0.015	-0.010	0.059	3.60
1PL Simulation - Random NR, Score Omit/NR=zero	0.037	-0.025	0.079	6.80
1PL Simulation - Systematic NR Score Omit/NR=blank	0.015	0.006	0.048	0.80
1PL Simulation -Systematic NR, Score Omit=zero, NR=blank	0.014	-0.012	0.052	1.90
1PL Simulation- Systematic NR Score Omit/NR=zero	0.027	-0.019	0.059	2.90
2PL Simulation - Random NR, Score Omit/NR=blank	0.014	0.007	0.055	1.50
2PL Simulation - Random NR, Score Omit=zero, NR=blank	0.014	-0.012	0.060	2.90
2PL Simulation - Random NR, Score Omit/NR=zero	0.039	-0.027	0.086	8.30
2PL Simulation - Systematic NR Score Omit/NR=blank	0.012	0.006	0.047	0.60
2PLSimulation-Systematic NR, Score Omit=zero, NR=blank	0.012	-0.012	0.054	2.00
2PL Simulation- Systematic NR Score Omit/NR=zero	0.024	-0.019	0.059	2.60

With respect to the direction of Bias, the "ignored/left blank" scoring treatment resulted in estimated abilities that were slightly higher than the true abilities. When omitted items and not-reached items were scored as incorrect with "0", the estimated abilities were less than the true abilities. This was also true of the scoring option where omitted items were scored as incorrect but not-reached items were left blank, however Bias was less extreme. Overall, Bias was in the same direction but larger in the calibrations where 20% of the examinees did not reach the end of the test relative to the 10% condition. These results

support Ludlow and Leary's (1999) findings concerning estimated abilities and the overall direction of Bias.

Table 3) Summary of Average RMSD, Bias, and Outlier Abilities Across 10 Data Simulations / WINSTEPS Executions for the "1.62% Omit Rate / 20% Not-Reached Condition"

	RMSD Items	Ability Bias	RMSD Abilities	# of Absolute Differences between Estimated and True Abilities >= 0.30
1PL Simulation - Random NR, Score Omit/NR=blank	0.025	0.013	0.081	5.10
1PL Simulation - Random NR, Score Omit=zero, NR=blank	0.023	-0.023	0.085	7.40
1PL Simulation - Random NR, Score Omit/NR=zero	0.074	-0.052	0.120	15.30
1PL Simulation - Systematic NR Score Omit/NR=blank	0.020	0.013	0.068	2.50
1PL Simulation -Systematic NR, Score Omit=zero, NR=blank	0.020	-0.023	0.075	4.40
1PL Simulation-Systematic NR Score Omit/NR=zero	0.043	-0.036	0.084	6.20
2PL Simulation - Random NR, Score Omit/NR=blank	0.024	0.014	0.076	4.60
2PL Simulation - Random NR, Score Omit=zero, NR=blank	0.023	-0.022	0.082	6.50
2PL Simulation - Random NR, Score Omit/NR=zero	0.074	-0.051	0.122	15.80
2PL Simulation - Systematic NR Score Omit/NR=blank	0.019	0.015	0.068	1.90
2PLSimulation-Systematic NR, Score Omit=zero, NR=blank	0.017	-0.021	0.073	3.80
2PL Simulation-Systematic NR Score Omit/NR=zero	0.047	-0.036	0.082	6.60

Across both the 10% and 20% not-reached conditions, the RMSD for abilities was lowest for the "ignore/leave blank" scoring option. This was followed by the scoring option where omits were scored as incorrect with 0 and not-reached left blank. It should

be noted that with the ".81% omit rate/10% not-reached' condition these two scoring treatments produced the same results when the data was simulated utilizing the one-parameter model and not reaching the end of the test was random or independent of ability. The RMSD for abilities was greatest and the performance poorest when omits and not-reached items were scored as incorrect with a zero.

The number of examinees whose absolute difference between their estimated and true ability was greater than or equal to 0.30 is reported in the last column of both Tables 2 and 3. Across both the ".81% omit rate/10% not-reached" and the "1.62% omit rate/ 20% not-reached" conditions, the "ignore/leave blank" scoring option performed best and produced the fewest "large/outlier" differences between the estimated and true abilities. This was followed by the scoring option where omits were scored as incorrect with 0 and not-reached left blank. The scoring option where omitted and not-reached items were scored as incorrect with 0 produced the greatest number of large/outlier differences between the true and estimated abilities. This poorer performance was more noticeable with larger amounts of missing data ("1.62% omit rate/ 20% not-reached "condition) and when not reaching the end of the test was independent of ability and determined through random selection.

The results of testing for significant differences in means across the ten runs for ability RMSDs are reported Table 4 and for ability differences greater than or equal to 0.30 in Table 5. Means and *P*-values are reported with the null hypothesis being that the means are statistically equal. *P*-values that are less than .05 are identified to indicate the rejection of the null hypothesis at the a =.05 level. With respect to the ability RMSDs reported in Table 4, eleven of the twelve differences in mean RMSDs were statistically

significant under both the "0.81% omit rate/ 10% not-reached" and the "1.62% omit rate/ 20% not-reached" conditions with the one exception being the insignificant difference between the "ignore/leave blank" and "omit=0 and not-reached=blank" scoring treatment when data was simulated with the 1PL and not-reached was assigned randomly.

Table 4) Dependent/Paired Difference T-Tests For Ability RMSDs Across 10 Runs

	Mean Ability RMSD			***P*-values For Paired Difference t-test Across 10 Runs**		
0.81% Omit/10% NR	**Omit & NR = Blank**	**Omit = 0 And NR=Blank**	**Omit And NR=0**	**Omit & NR = Blank w/ Omit = 0 & NR=Blank**	**Omit &NR =Blank w/ Omit & NR=0**	**Omit = 0 & NR=Blank w/ Omit & NR=0**
1PL Random	.059	.059	.079	.926	.002*****	.000*****
1PL Systematic	.048	.052	.059	.028*****	.001*****	.000*****
2PL Random	.055	.060	.086	.044*****	.000*****	.000*****
2PL Systematic	.047	.054	.059	.030*****	.001*****	.001*****
1.62% Omit/20% NR						
1PL Random	.081	.085	.120	.306	.000*****	.000*****
1PL Systematic	.068	.075	.084	.000*****	.000*****	.000*****
2PL Random	.076	.082	.122	.003*****	.000*****	.000*****
2PL Systematic	.068	.073	.082	.001*	.000*	.000*

Note: * indicates significant difference (two-tailed) at the *a* = .05 level.

For the ability differences greater than or equal to 0.30 reported in Table 5, nine of the twelve differences in the means were statistically significant under the "0.81% omit rate/ 10% not-reached" condition. The three exceptions were the insignificant differences between the "ignore/leave blank" and "omit=0 and not-reached=blank" scoring treatments when data was simulated with the 1PL and the difference between the "omit and not-reached=0" and the "omit=0 and not-reached=blank" scoring treatments when

data was simulated with the 2PL and not-reached was assigned systematically. All twelve of the differences in means were statistically significant with larger amounts of missing data as reported under the "1.62% omit rate/ 20% not-reached" condition.

Table 5) Dependent/Paired Difference T-Tests For Ability Differences >= 0.30 Across 10 Runs

	Mean # of Persons w/ Ability Differences >= 0.30			***P*-values For Paired Difference t-test Across 10 Runs**		
0.81% Omit/10% NR	**Omit & NR = Blank**	**Omit = 0 And NR=Blank**	**Omit And NR=0**	**Omit & NR = Blank w/ Omit = 0 & NR=Blank**	**Omit &NR =Blank w/ Omit & NR=0**	**Omit = 0 & NR=Blank w/ Omit & NR=0**
1PL Random	3.20	3.60	6.80	.494	.008*****	.005*****
1PL Systematic	0.80	1.90	2.90	.066	.004*****	.004*****
2PL Random	1.50	2.90	8.30	.013*****	.000*****	.000*****
2PL Systematic	0.60	2.00	2.60	.010*****	.006*****	.051
1.62% Omit/20% NR						
1PL Random	5.10	7.40	15.30	.012*****	.000*****	.000*****
1PL Systematic	2.50	4.40	6.20	.030*****	.002*****	.003*****
2PL Random	4.60	6.50	15.80	.022*****	.000*****	.000*****
2PL Systematic	1.90	3.80	6.60	.002*****	.000*****	.003*****

Note: * indicates significant difference (two-tailed) at the *a* = .05 level.

Discussion

In terms of Bias, RMSD and the number of examinee ability estimates that differed from their "true" ability estimates by 0.30 or more, the scoring treatment that consistently produced the best results was when omitted and not-reached items were ignored or left blank. This scoring option yielded the lowest absolute Bias, lowest RMSD for abilities and the lowest number of outlier examinee abilities across both the ".81% omit rate/ 10%

not-reached" and "1.62% omit rate/20% not-reached" conditions. In addition the performance of the "ignore/leave blank" scoring option seemed to improve relative to the other two scoring treatments as the amount of missing data increased. This stands in contrast to the scoring option where omits and not reached were scored as incorrect with a "0". This scoring treatment yielded the poorest results from both a statistical and practical perspective and the performance seemed to worsen relative to the other two scoring treatments as the amount of missing data increased.

One of the contributing factors to the comparative strength of the "ignore/leave blank" scoring treatment lies with the estimation method itself. With JMLE, an item's difficulty estimate is based upon only those examinees who responded to that item which is in contrast to an estimation method such as MMLE where item difficulty estimates are based upon an assumed ability distribution. In addition, JMLE estimates item difficulties and examinee abilities simultaneously and as with the item difficulty estimates, examinee abilities are based solely on those items that the examinee responds to. This finding supports previous findings (Linacre 1999, Demars 2003 and Shin 2009) that Joint Maximum Likelihood Estimation is robust with missing data.

It was also evident that across all scoring treatments, performance in terms of ability RMSDs and outlier examinee abilities, was poorest when not reaching the end of the test was random or independent of ability. The results across all scoring options improved and became more similar when not reaching the end of the test was systematic or dependent upon ability. The link between examinee ability and omitting behavior and not-reaching the end of a test is one that requires additional study and is an important

consideration for practitioners when selecting the appropriate scoring treatment for omitted and not reached-items.

The finding of significant differences between the "ignore/leave blank" and the "omit and not-reached = 0" scoring treatments is important. Though all of the differences were statistically significant, practical significance between these scoring options was also evident with larger amounts of missing data and when not-reaching the end of the test was independent of ability. The practice of treating missing data as incorrect is common with the large-scale assessments often associated with state testing programs. However, based on the results of the current study this would result in the greatest number of differences between students' true and estimated abilities and as a practice should be approached with caution.

Lastly and in contrast to the above, the differences between the "ignore/leave blank" and the "omit=0/not-reached =blank" scoring treatments were less pronounced. Though statistically significant the differences between these scoring options were of somewhat questionable practical significance and it is this issue of practical significance that may carry more weight given the circumstances and constraints that practitioners encounter.

Limitations

This study utilized the Rasch model and Joint Maximum Likelihood Estimation with fixed values for key variables such as sample size, the number of items and the level of omitted and not-reached items. In addition, a strong assumption was made regarding the sequencing of items from least to most difficult during the simulation process. The

results should in turn be interpreted cautiously not only in light of the model chosen and estimation method used but also in light of the controls and constraints placed upon the above variables.

The three scoring treatments for omitted and not-reached items that have been the focus of this study have appeal to practitioners because of their ease of implementation and the uniform manner in which they can be applied. However, there exist other and more sophisticated methods for imputing item scores for omitted and not-reached items that though beyond the scope of this paper deserve to be considered and evaluated.

References

DeAyala, R.J., Plake, B.S, Impara, J.C., Kozmicky, Michelle (2000). *The Effect of Omitted Responses on Ability Estimation in IRT.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA., April 24-28, 2000. [ERIC Document Reproduction Service No. ED 447 167]

DeMars, C. (2003) *Missing Data and IRT Item Parameter Estimation.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL., April , 2003. [ERIC Document Reproduction Service No. ED 476 175]

Han, K. T. (2007). *WinGen2: Windows software that generates IRT parameters and item responses* [computer program]. Amherst, MA: University of Massachusetts, Center for Educational Assessment. Retrieved May 13, 2007, from http://www.umass.edu/remp/software/wingen/

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1993) *Omitted and Not-Reached Items in Mathematics in the 1990 National Assessment of Educational Progress CSE Technical Report 357*. Center for Research on Standards, and Student Testing, Los Angeles, CA: [ERIC Document Reproduction Service No. ED 378 220]

Linacre, J. M. (1999). Understanding Rasch Measurement: Estimation Methods For Rasch Measures. *Journal of Outcome Measurement, 3,* 382-405.

Linacre, J.M (1991-2002). *WINSTEPS 3.65* [Computer Software]. Chicago, IL.: John M. Linacre www.WINSTEPS.com.

Linacre, J.M (1991-2002). *A User's Guide to WINSTEPS MINISTEP Rasch-Model Computer Program*. Chicago, IL.: Winsteps.com.

Lord, F.M. (1974) Estimation of Latent Ability and Item Parameters When There Are Omitted Responses. *Psychometrika*, 39, 247-264.

Lord, F.M. (1980) *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Ludlow, L.H. & O'Leary, M. (1999). Scoring Omitted and Not-Reached Items: Practical Data Analysis Implications. *Educational and Psychological Measurement*, 59, 615-630.

Mislevy, R.J., & Wu, P.K. (1988) *Inferring Examinee Ability When Some Item Responses Are Missing.* Princeton, NJ: Educational Testing Service. [ERIC Document Reproduction Service No. ED 395 017].

Shin, Seon-HI. (2009). How to Treat Omitted Responses in Rasch Model-Based Equating. *Practical Assessment, Research & Evaluation,* Volume 14, Number 1.

Appendix A

A.1) Schedule of Item Omits for the "0.81% omit rate - 10% Not-Reached Condition "

Item Number	Number of Examinees Omitting Each Item
1	0
2-3	1
4-6	2
7-14	3
15-23	4
24-32	5
33-40	6

*****Example of how to read table A.1: Item 1 was not omitted by anyone. Each of items 15-23 was omitted 4 times by randomly selected examinees.

A.2) Schedule of Not-Reached for Items 33-40 for the "0.81% omit rate - 10% Not-Reached Condition"

Item Response Strings With Missing/Blank Fill Prior To Scoring	Number Of Examinees Not Attempting the Response String
33-40	3
34-40	4
35-40	5
36-40	6
37-40	7
38-40	8
39-40	9
40	10
Total w/ Percent	52 (10.4%)

*****Example of how to read table A.2: Three examinees were selected either randomly or systematically (see A.5 for systematic selection criteria) to set the item response string made up of items 33-40 to missing. The last item that these 3 examinees reached was item 32. Likewise, 7 examinees were selected either randomly or systematically to set the item response string made up of items 37-40 to missing. A total of 52 examinees (10.4%) of the total did not reach the end of the test.

A.3) Schedule of Item Omits for the "1.62% omit rate - 20% Not-Reached Condition"

Item Number	Number of Examinees Omitting Each Item
1	1
2-3	2
4-6	4
7-14	6
15-23	8
24-32	10
33-40	12

*Example of how to read table A.3: Item 1 was omitted by 1randomly selected examinee. Each of items 15-23 was omitted 8 times by randomly selected examinees.

A.4) Schedule of Not-Reached For Items 33-40 for the "1.62% omit rate - 20% Not-Reached Condition "

Item Response Strings With Missing/Blank Fill Prior To Scoring	Number Of Examinees Not Attempting the Response String
33-40	6
34-40	8
35-40	10
36-40	12
37-40	14
38-40	16
39-40	18
40	20
Total w/ Percent	104 (20.8%)

*****Example of how to read table A.4: Six examinees were selected either randomly or systematically (see A.5 for systematic selection criteria) to set the item response string made up of items 33-40 to missing. The last item that these 6 examinees reached was item 32. Likewise, 14 examinees were selected either randomly or systematically to set the item response string made up of items 37-40 to missing. A total of 104 examinees (20.8%) of the total did not reach the end of the test.

A.5) Schedule of Examinee Selection for Not-Reaching the End of the Test When Not-Reached Is Dependent on Ability (Systematic Condition)

	Examinees Available To Be Selected if "True" Ability (Theta) <= Upper Limit	
Not-Reached Item Response Sets	1PL Simulated Data "True" Ability (Theta) Upper Limit	2PL Simulated Data "True" Ability (Theta) Upper Limit
33-40, 34-40	-.87	-.78
35-40, 36-40	-.26	-.24
37-40	.23	.29
38-40	.83	.80
39-40, 40	No limit - anyone	No limit - anyone

*****Example of how to read table A.5: Only the lowest performing examinees, those with a true ability (theta) of less than or equal to -.87 were available to be randomly selected to not reach item set 33-40 or item set 34-40. Any examinee with a theta less than or equal to -.26 including those in the above group were available to be selected to not reach item set 35-40 or item set 36-40. Any examinee could be selected to not reach item set 39-40 or item 40.