Practical Assessment, Research, and Evaluation

Volume 14 *Volume 14, 2009* Article 1

November 2019

How to Treat Omitted Responses in Rasch Model-Based Equating

Seon-Hi Shin

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 14, Number 1, January 2009 ISSN 1531-7714

How to Treat Omitted Responses in Rasch Model-Based Equating

Seon-Hi Shin, *California State University, Long Beach*

This study investigated the impact of the coding scheme on IRT-based true score equating under a common-item nonequivalent groups design. Two different coding schemes under investigation were carried out by assigning either a zero or a blank to a missing item response in the equating data. The investigation involved a comparison study using actual large scale data and then Monte Carlo simulations for a systematic inspection on the topic. The recommendations on the basis of the findings of the study were made to treat omitted responses as not-administered rather than as wrong, and use a large sample size to ensure the accuracy of the screening tools such as the displacement index and the robust-z statistic during equating.

In the literature of psychometrics, the impact of treating omitted responses as incorrect on the estimation of ability and/or item parameters in the item response theory (IRT) context has been reported over years. Several researchers (Lord, 1974, 1983; Mislevy & Wu, 1996) used modeling techniques to handle omissions when estimating individual ability/item parameters directly while other researchers (Ayala, Plake, & Impara, 2001; Ludlow & O'Leary, 1999) compared different estimation strategies for handling missing data. These studies generally agree that treating omitted responses as if they were wrong is not appropriate. Their arguments were based on the results from estimating individual subjects' latent ability (θ) directly. However, in actual large scale testing situations, individual students' theta scores are determined by mapping raw scores to theta scores in the conversion table obtained from equating instead of estimating students' thetas directly. Few empirical studies have been reported with regard to the impact of the coding scheme of omitted responses on IRT-based true score equating which is frequently used in large scale testing field settings. Specifically, a common-item nonequivalent groups design (Kolen & Brennan, 2004) has often been employed in conjunction with one or more IRT models in the operational equating field. Under the design, common items are treated as anchor or linking items which play a crucial role to link different test forms of a test to maintain the scale integrity. Among alternatives, the Rasch model (Rasch, 1960) for dichotomous items and the Partial Credit Model (Masters, 1982) for polytomous items are often fitted in K – 12 standardized achievement testing.

This study investigated the impact of the coding scheme on IRT-based true score equating under a common-item nonequivalent groups design. Two different coding schemes under investigation were carried out by assigning either a zero or a blank to a missing item response in the equating data. Treating missing as incorrect instead of blank results in different item statistics such as item difficulty and discrimination defined under the classical test theory framework. The present study, in particular, centers on illustrating the effects of such different treatments of omitted responses on detecting drift anchor items during Rasch model-based equating. The calibration computer software used was WINSTEPS (Version 3.63.2).

The reliability and validity of equating results are substantially affected by the process called screening anchor items. The screening process involves comparing the fixed values (typically, coming from the item bank for the test or from the base form) with the estimated parameters using current equating data for anchor items. The item significantly drifting from the fixed value should

be dropped from the final anchor set. The following two psychometric tools are frequently used as criteria for detecting drift items when the common item parameters are fixed to the values from the base form at calibration:

Displacement (D_i) index approximates the deviation of the item difficulty estimate ($\hat{b}_i$) for item i from the statistically better value ($\bar{b}_i$) which would result from the best fit of the given data to the model (Linacre, 2005):

$$D_i = \frac{(\hat{b}_i - \bar{b}_i)}{\sigma_i^2} \quad (1)$$

where σ_i^2 denotes the model-derived variance of the item difficulty estimate. A valid value for the displacement will be produced by WINSTEPS only for the anchor items by fixing the anchor item parameters to given values at calibration. The cut points of the displacement regarded as significant drifting are typically ± 0.3, ± 0.4, or ± 0.5 in the literature (Miller, Rotou, & Twing, 2004; Shin, Lee, & Young, 2007).

Meanwhile, the robust-z statistic (Tenenbaum, Lindsay, Siskind, Wall-Mitchell, & Saunders, 2001) requires fixing anchor item parameters to given values at one calibration and setting them free to be estimated at another calibration:

$$Z_i = \frac{[(b_{iF} - \hat{b}_{iE}) - M_d]}{INQ_d * 0.74} \quad (2)$$

where b_{iF} stands for the fixed value, and $\hat{b}_{iE}$ denotes the estimated item difficulty for item i. The median and inter-quartile range of the differences between the fixed values and the item difficulty estimates across all the items in the anchor set are represented by M_d and INQ_d, respectively. If the absolute value of the robust-z for an item is equal to or larger than 1.96, the item is typically flagged as drifting.

One can argue that if the pattern and the total number of the items flagged by these diagnostic tools differ substantially between different coding schemes for omitted responses, the resulting raw-to-theta score conversion table will more likely differ. Subsequently, individual students will more likely earn different theta scores. Therefore, it is worthwhile to address the relationship between the coding scheme and the performance of these anchor item screening tools empirically when missing data are present. The current article included a comparison study using actual large scale data first and then Monte Carlo simulations for a more systematic and comprehensive inspection on the topic.

A COMPARISON STUDY

A total of 2,941 students' response strings in a standardized English Language Arts (ELA) test were used in this study. The test was administered to grade three students in a large school district in summer, 2006. Fifty multiple choice items composed the test. The raw score was obtained by summing individual dichotomous item scores. The psychometric literature has reported that the act of omission is related to the examinee's ability. To look at such a relationship, the students' raw scores were grouped into one of the three ability categories: high, medium, low. SAS PROC RANK (Version 9.1) was used in producing the ranks. The average number of missing items for each ability group from the lowest to the highest was 4.5, 1.5, and 0.8, respectively. Although the test was not a speed test, more students omitted their responses to the items toward the end of the test, regardless of ability (Figure 1). For instance, 20.6% of low ability group, 10.1% of medium ability group, and 7.6% of high ability group omitted their responses to the last item in the test.

All test items were selected via the stand-alone field test equating conducted prior to the operational administration, and thus all the items of the test were regarded as anchor items at the start of equating. The students' omitted item responses were coded either as wrong (zero hereafter) or as missing. The option, MISSCORE = -1 was included in the control file of WINSTEPS (Version 3.63.2) to ignore omitted responses for the missing condition. The robust-z is much more influenced by the presence of the other anchor items in the computation when evaluating an anchor item (Refer to equation (2) for the reason). Therefore, for each condition, the anchor items were examined for drift by the displacement first until no drift item was left by this index. Then, the robust-z was applied to the remaining items in the anchor set. Items were flagged as drifting either if the absolute value of the displacement was larger than 0.4 or if the absolute value of the robust-z was equal to or larger than 1.96. The zero condition showed 25 drift items while the missing condition resulted in 26 drift items by the displacement criterion, respectively. In both coding conditions, after the items identified as drifting by the displacement were removed from the anchor set, no additional items were flagged by applying the robust-z criterion. The items flagged in the zero condition were also flagged in the missing condition except one item. Additionally, two new items were identified as drifting in the missing condition. In both coding conditions anchor items accounted for about 50% of the test length. This percentage was much higher than approximately 20% of the total number of test items recommended by Angoff (1971) as the minimum number of anchor items in common-item linking. Note that the percentage of the

anchor items in the test length is typically about 40% or higher in large scale testing practice.

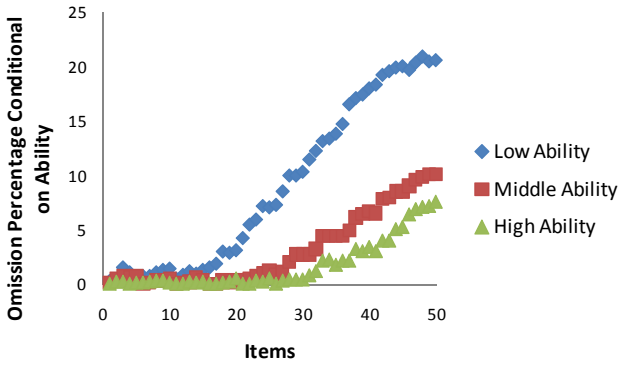

Figure 1. Percentage of Omitted Responses to Each Item Conditional on Ability

The raw-to-theta score conversion tables were obtained as part of the results of equating for the two coding conditions. The differences between the thetas of both conditions ranged from 0.01 to 0.03 across the fifty different raw score points. Both conditions showed almost identical standard errors of estimates for individual thetas. The largest difference between the two conditions was only 0.0008. Accordingly, the 95% confidence interval (C.I.) was constructed by multiplying the standard errors of estimates of the missing condition by ±1.96 to evaluate the significance of the difference between the thetas of both conditions. Figure 2 portraits the C.I. band along with the theta difference within it against raw score points. The difference was ignorable. The scale score was simply a linear combination of the theta score, which had a one-to-one mapping relationship with the raw score. Consequently, the difference between the scale scores of both conditions should remain the same as shown in Figure 2 except the measurement units on the Y axis of the plot.

The findings of this comparison study indicated that the impact of the coding scheme on equating was ignorable given the observed degree of omitted responses. The two coding schemes resulted in almost the same anchor set at equating, and as a result produced almost identical raw-to-theta score conversion tables. Therefore, it could be argued that equating was robust against different coding schemes for omitted responses. In order to investigate this phenomenon in a more systematic and comprehensive way, Monte Carlo simulations were conducted. The results of the simulation study are illustrated in the following section.

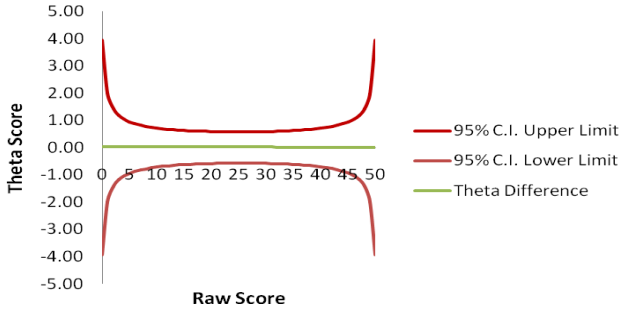

Figure 2. Theta Difference between Missing and Zero Coding Schemes within the 95% Confidence Interval

MONTE CARLO SIMULATIONS

Item Parameters

Fifty multiple choice item parameters were obtained from two years (2005 and 2006) of equating results for the ELA test administered to seventh graders from the same school district. Out of the item parameter estimates calibrated through the 2005 equating, twenty anchor items were borrowed for the study. The difficulty of the chosen anchor items ranged from easy to hard. Among the twenty anchor items, five drift items were created by adding +1.2 logits to the borrowed estimates. This logit value was chosen because a previous Monte Carlo study (Shin, Lee, & Young, 2007) reported that the drift direction caused little difference in detecting true drift items. This chosen drift magnitude was large enough to be detected by the screening statistics. The thirty non-anchor item parameters were borrowed from the 2006 equating results. The anchor items were positioned before the non-anchor items in the test. The location and the parameter values of the anchor items for the study are shown in Table 1.

Data Generation

Different levels of omitted responses were generated for different ability groups to mimic the actual missing data characteristics observed in the comparison study. The population distribution of ability was assumed to be the standard normal. The theta scores of individual students were randomly sampled using the SAS built-in standard normal random number generator. Based on the generated theta scores, three ability groups were formed using SAS PROC RANK. Omissions occurred either in the last ten or five non-anchor items or in the ten or five anchor items including the drift items (Table 1). The drift items approximately spanned the difficulty range for the five anchor item omission condition, and the item right after each drift item was additionally chosen for the ten anchor item omission condition. Complete data sets were also

created to set the baseline performance with which to compare the results of missing and zero conditions. The conditional percentages of omitted responses were 7%, 10%, 20% for high, medium, and low ability groups, respectively. The sample size was manipulated to be N = 200, 500, 1000, and 3000. The item response string for each theta score in each of the experimental conditions was created using the item parameters, assuming the Rasch model as the true IRT model.

Table 1. Parameters of Drift and Non-drift Anchor Items

Item	No drift	Drift to +1.2 logits
1	-1.7514	-1.7514
2[DO]	-1.3795	-0.1795
3[O]	-1.3171	-1.3171
4[DO]	-1.0149	0.1851
5[O]	-0.7269	-0.7269
6	-0.5440	-0.5440
7	-0.3833	-0.3833
8	-0.2472	-0.2472
9[DO]	-0.0725	1.1275
10[O]	0.0909	0.0909
11	0.1125	0.1125
12	0.1399	0.1399
13	0.2388	0.2388
14	0.3960	0.3960
15	0.4599	0.4599
16[DO]	0.7333	1.9333
17[O]	1.0146	1.0146
18	1.1087	1.1087
19[DO]	1.2065	2.4065
20[O]	1.7798	1.7798

[D] Drift item
[O] Omitted responses

Measures

Therefore, there were a total of forty eight conditions. For each condition the displacement and robust-z were computed. There were 100 replications for each condition. The frequency of flagging drift items correctly was defined as hit frequency for each run whereas the frequency of flagging non-drift items falsely was defined as false frequency. Then, the averages of the hit and the false frequencies over replications were computed for each of the forty eight conditions. Hence, there were a pair of average hit frequencies and also a pair of average false frequencies according to the anchor item screening statistic used (i.e., displacement and robust-z), for each condition. These average statistics were the primary measures for this study.

RESULTS

When omissions occurred in non-anchor items, the difference in the average hit frequencies of the two coding schemes was ignorable, regardless of the sample size, the anchor item screening statistic, and the number of the items having omitted responses. The largest difference was 0.04 and occurred in the condition where the number of the non-anchor items having omitted responses was ten, the sample size was 200, and the screening statistic used was the robust-z. Similarly, the difference between the average false frequencies of the two coding schemes was ignorable when omissions occurred in non-anchor items. In addition, the results from the complete data sets appeared almost identical to those from the missing and the zero coding schemes in the non-anchor item omission conditions (Figures 3 and 4).

In contrast, when the anchor items including the drift items were associated with omitted responses, the difference between the two coding schemes became evident in the average false frequency by the displacement and in the average hit frequency by the robust-z, respectively. For instance, the zero conditions flagged many more items falsely in using the displacement when five anchor items had omitted responses. The average false frequency, though, decreased as the sample size increased for both coding schemes. However, when ten anchor items were associated with omitted responses, the difference between the two coding schemes in the average false frequency became even greater. Furthermore, the average false frequency of the zero coding condition did not decrease as the sample size increased (Figure 3). For example, the zero coding flagged an average of 4.6 non-anchor items falsely while the missing coding flagged 1.1 non-anchor items incorrectly on average at the sample size of 3000. When omissions were accompanied by the missing coding, the average false frequency decreased dramatically as the sample size increased while holing the average hit frequency perfect (i.e., all five true drift items were correctly detected in all replications). In fact, the average hit frequency by the displacement appeared approximately saturated in both coding schemes in all the studied experimental conditions. In other words, the displacement flagged not only the five true drift items correctly but also non-drift items falsely in all conditions to different degrees.

Practical Assessment, Research & Evaluation, Vol 14, No 1
Shin, Omits in Rasch Equating

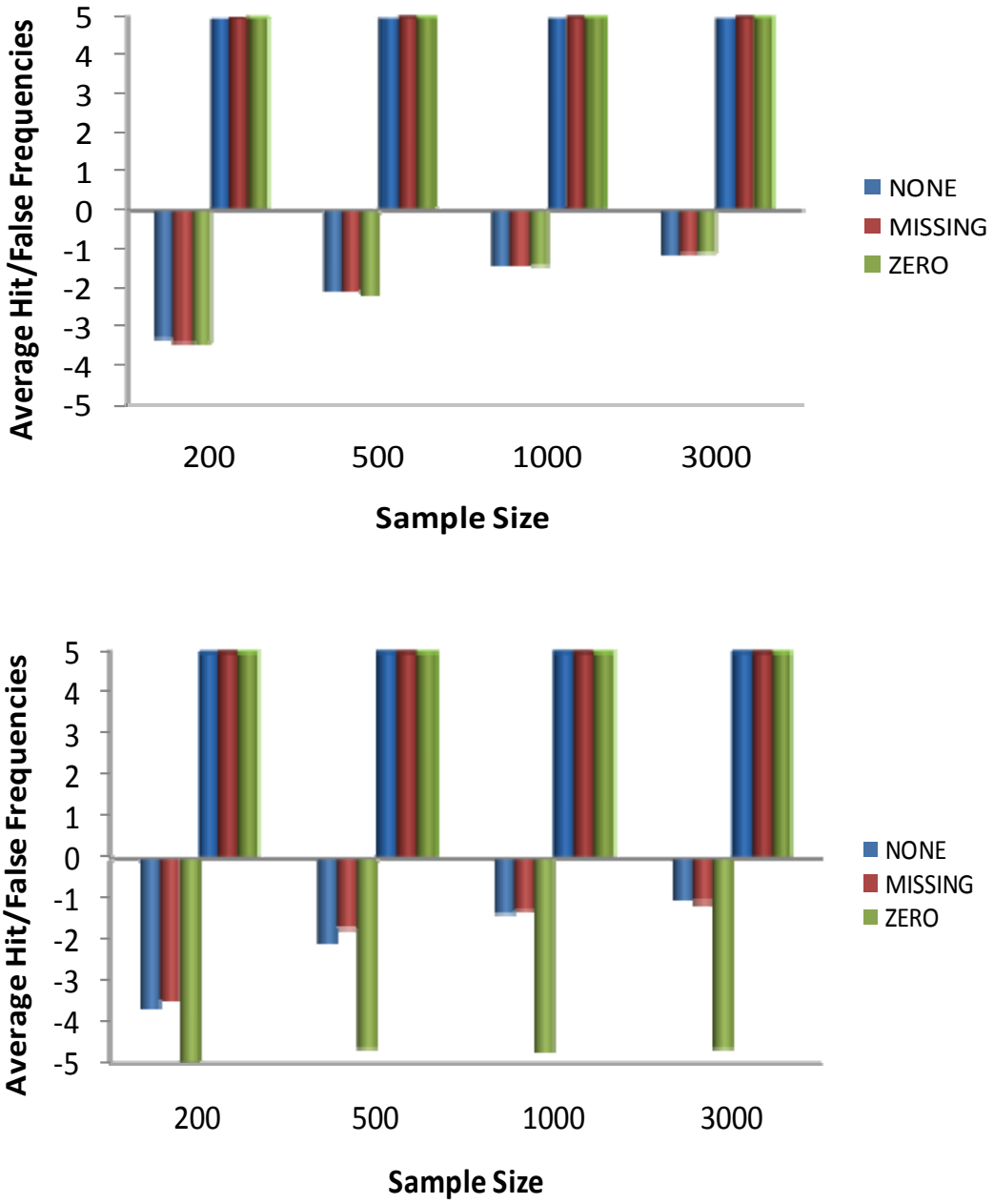

Figure 3. Average Hit (positive on the ordinate) and False (negative on the ordinate) Frequencies by Displacement when Omitted Responses Occurred in Ten Non-anchor Items (Top) and Ten Anchor Items (Bottom)

Note. NONE represents the condition of no omitted response.

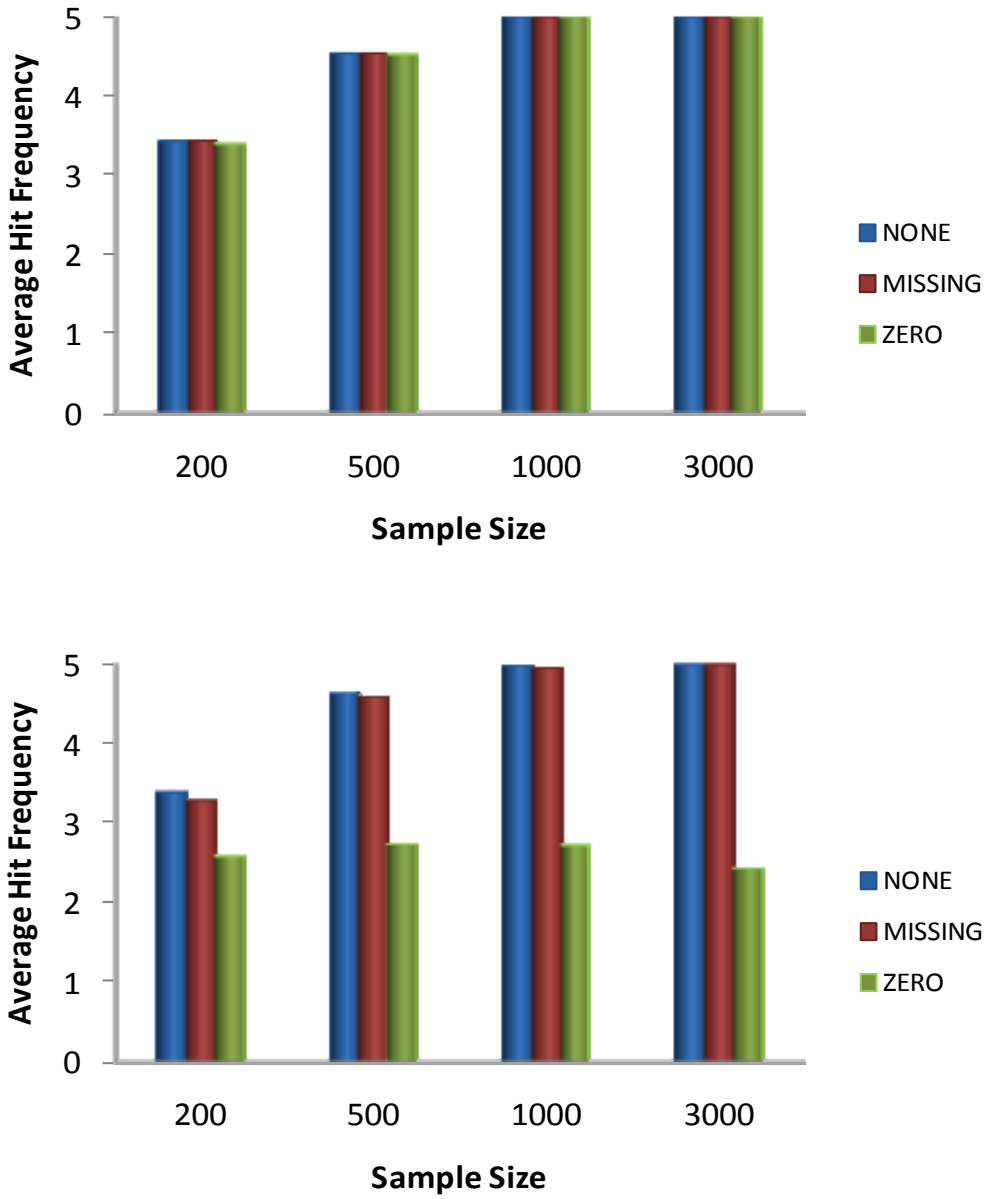

Figure 4. Average Hit Frequency by Robust-*Z* when Omitted Responses Occurred in Ten Non-anchor Items (Top) and Ten Anchor Items (Bottom)

Note. NONE represents the condition of no omitted response.

Meanwhile, the average hit frequency by the robust-z of the missing coding grew from 3.29 to 5.0 as the sample size increased from $N = 200$ to $N = 3000$ when omissions occurred in the ten anchor items (Figure 4). However, the average hit frequency of the zero coding over the same sample size range ranged from 2.43 to 2.74. Additionally, there was no monotonic increase in the average hit frequency over the increase of the sample size for this coding. Interestingly, there was only a small or little difference between the two coding schemes in the average hit frequency by the robust-z for the conditions where omissions occurred in the five anchor items. The robust-z did not flag non-anchor items falsely in any condition.

DISCUSSION

Equating is a crucial psychometric procedure in large scale testing. The equating procedure involves the screening process where drift items in the anchor set are identified. Specifically, for the Rasch model-based true score equating, the screening statistics such as the displacement and the robust-z are frequently utilized at the item level screening. Using an adequate set of anchor items during equating is directly related to the success of equating to maintain the integrity of the existing scale.

Considering the importance of the anchor set, omitted responses in the equating data should be treated with caution. Omitted responses are often treated as either wrong or ignorable (equivalently, not-administered). The findings of the comparison study presented in the earlier section of this article indicated that the two coding schemes under consideration made little difference in flagging drift items, and subsequently in the raw-to-scale score conversion table. However, it needs to be noted that one cannot know which items are true drift items when analyzing actual testing data.

The systematic inspection through Monte Carlo simulations, on the other hand, uncovered that the two coding schemes made little difference in detecting true drift items as long as omitted responses occurred in non-anchor items. This finding per se was not surprising because the displacement and the robust-z are defined for anchor items only, although the estimation of the parameters of both anchor and non-anchor items is completed simultaneously in one WINSTEPS run.

On the contrary, when omitted responses occurred in anchor items including drift items, the missing coding outperformed the zero coding substantially, and its outperformance improved monotonically as the sample size increased. This phenomenon was particularly true with the displacement. When the robust-z was used as the screening tool, the outperformance of the missing coding was obvious only when omissions were associated with ten anchor items. In this omission condition, the missing coding flagged an average of 5.0 drift items correctly at the sample size of 3000 whereas the zero coding flagged only 2.4 drift items correctly on average. Apparently, the impact of the coding scheme became more evident as more anchor items including true drift items were associated with omissions. Thus, placing anchor items in the front part of the test may help reduce omitted responses to the anchor items in practice.

In general, the displacement appeared hyperactive in the sense that it flagged non-drift items in addition to the true drift items whereas the robust-z was relatively inactive and hence identified all of or fewer than the true drift items without placing a flag on a non-drift item. It was also clear that increasing the sample size improved the accuracy of the screening tools even when the data were complete or when omissions did not occur in the anchor items. The displacement required a larger sample than the robust-z to reach the plateau of the perfect performance. However, when many anchor items (e.g., ten anchor items including drift items) were associated with omissions, the increase of the sample size did not necessarily improve the performance of the screening statistics when the omitted responses were treated as wrong.

Taken together, the findings of the present study provide useful guidelines for psychometric practitioners in large scale testing field settings, although they can be generalized only to the conditions similar to those investigated in the study. The study suggests that one leave omitted responses as missing (namely, treat them as not-administered), and use a large sample size to ensure the accuracy of the screening tools during equating. Fortunately, using a large sample does not impose a burden in large scale testing. The current study also illustrated the usefulness of conducting Monte Carlo simulations. The findings from the comparison study where the true parameters were hardly known failed to show a full picture but only illustrated a limited snapshot, which could be misleading. However, the findings from the simulation study unveiled various facets of the impact of the coding scheme on equating.

Note that the screening statistics investigated in the present study are appropriate specifically for the type of true score equating which requires fixing common item parameters to the values from the base form at calibration. The impacts of the two different coding schemes on the other type of true score equating which requires separate estimation across forms cannot be inferred from the results of this study. Therefore, further research is needed to provide guidelines for how to treat omitted responses in equating accompanied with separate item parameter estimation procedures such as item characteristic curve methods (Stocking-Lord and Haebara) and two moment methods (Mean/Mean and Mean/Sigma). Different IRT models also need to be included in the further research.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education

Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement,* 38(3), 213 − 234.

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer-Verlag.

Linacre, J. M. (2005). *A user's guide to WINSTEPS MINISTEP Rasch-model computer programs*. Chicago, IL: Winsteps.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, 39, 247 − 264.

Lord, F. M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, 48, 477 − 481.

Ludlow, L. H., & O'Leary M. (1999). Scoring omitted and not-reached items: practical data analysis implications. *Educational and Psychological Measurement*, 59(4), 615 − 630.

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

Miller, G., Rotou, O., & Twing, J. (2004). Evaluation of the 0.3 logits screening criterion in common item equating. *Journal of Applied Measurement*, 5, 172 – 177.

Mislevy, R. J., &Wu, P-K. (1996). Missing responses and IRT ability estimation: omits, choice, time limits, and adaptive testing (Research Report No. RR-96-30-ONR). Princeton, NJ: Educational Testing Service.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Shin, S., Lee, Y., & Young, M. J. (2007, April). Examining linking item diagnostic tools with noisy data. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois.

Tenenbaum, I., Lindsay, S., Siskind, T., Wall-Mitchell, M. E., & Saunders, J. (2001). Technical documentation for the 2000 Palmetto achievement challenge tests of English language arts and mathematics. Columbia, SC: South Carolina Department of Education.

Note

This research is the revised version of a study presented at the annual meetings of the American Educational Research Association, March, 2008, New York, New York.

Citation

Shin, Seon-Hi (2009). How to Treat Omitted Responses in Rasch Model-Based Equating. *Practical Assessment Research & Evaluation*, 14(1). Available online: http://pareonline.net/getvn.asp?v=14&n=1

Author

Seon-Hi Shin
Department of Educational Psychology, Administration and Counseling
College of Education, ED1-57
California State University, Long Beach
1250 Bellflower Blvd.
Long Beach, CA 90840-2201

E-mail: sshin [at] csulb.edu

Phone: (562) 985-2428
Fax: (562) 985-2428