4-2003

Missing Data and IRT Item Parameter Estimation

Christine E. DeMars
James Madison University, demarsce@jmu.edu

Missing Data and IRT Item Parameter Estimation

Christine DeMars

James Madison University

Paper presented at the annual meeting of the American Educational Research Association, Chicago. (2003, April).

Abstract

Non-randomly missing data has theoretically different implications for item parameter estimation depending on whether joint maximum likelihood or marginal maximum likelihood methods are used in the estimation. The objective of this paper is to illustrate what potentially can happen, under these estimation procedures, when there is an association between ability and the absence of response. In this example, data is missing because some students, particularly low-ability students, did not complete the test.

Missing Data and IRT Item Parameter Estimation

Data may be missing from a data set used for item calibration either by design (not all examinees were administered all items) or because examinees left some items unanswered. When the data are missing by design and not because of unanswered items, the design can be taken into account when calibrating the item parameters. For example, in the case of vertical equating the difficulty of the test form is to some degree matched to the ability of the examinees; students in a lower grade in school may be given an easier form of a math test than students in the next grade, with some middle-difficulty anchor items common to both forms. When the items are calibrated by joint maximum likelihood estimation (JMLE) techniques, it theoretically makes no difference that the non-anchor items were answered by groups of different ability and all items can be calibrated simultaneously (Hambleton & Swawinathan, 1985, p. 212; Hambleton, Swawinathan, & Rogers, 1991, p. 135; Lord, 1980, p. 200-201). If marginal maximum likelihood (MML) techniques are used for the estimation, the vertical equating design can be taken into account by estimating a separate trait distribution for each group of examinees (DeMars, 2002; Zimowski, Muraki, Mislevy, & Bock, 1996). If this procedure is not followed, DeMars (2002) showed that when MML estimation was used with a single latent ability distribution the items unique to the easier form (given only to the lower-ability students) had positively biased difficulties and the items unique to the more difficult form had negatively biased difficulties (as follows from MML theory).

A different situation emerges when the data are missing because some examinees do not answer all items. This may present more problems for MML because there are not identified groups for which separate latent ability distributions can be estimated as would be done with vertical equating. When items are omitted in the middle of a test, most sources agree the omitted

items should ***not*** be treated as if they had never been presented to the examinee (Mislevy & Wu, 1988; Lord 1974; Lord, 1980; Lord, 1983; Ludlow & O'Leary, 1999); generally such items are treated as incorrect or fractionally correct. Recently, De Ayala, Plake, and Impara's work (2001) showed more accurate theta estimation when omitted items were treated as half correct, compared to correct proportional to the number of alternatives, incorrect, or not-administered, but they did not examine recovery of item parameters.

In another situation which is the focus of the present study, some students do not have time to finish a test that is intended as a power (non-speeded) test. Some researchers treat the *not-reached* items at the end of the test as if they were *not-administered* items, as if examinees reaching different points in the test in effect were administered different forms (with all the preceding items as anchor items). Research has also addressed this procedure, which is conceptually similar to treating each set of reached items as a separate form and equating all forms in a simultaneous item calibration.

Lord (1980, p. 226) proposed that "If most examinees read and respond to items in serial order, a practical procedure for formula-scored tests is to ignore the 'not-reached' responses of each examinee when making statistical inferences about examinee and item parameters". As Lord noted, the assumption that examinees answer items in order is somewhat problematic. This was similar to another piece, where Lord suggested items not reached at the end of the test due to time limits could be ignored during the estimation (of item and examinee parameters). He explained "a key property of icc theory is that item parameters do not depend on the group of examinees tested, within reasonable limits; and that examinee ability ( ) does not depend on the items administered" (1974, p. 248).

Ludlow and O'Leary (1999) noted that treating not-reached items as incorrect can cause those items to have higher difficulties than they should have, as well as inflate goodness-of-fit statistics. They suggested treating not-reached items as missing during the item calibration, but as incorrect during the people calibration. Their focus, though, was on the people estimate; they showed how treating not-reached items as missing can lead to very different ability estimates than those obtained from treating not-reached items as incorrect.

Oshima (1994), using MML, explored the effects of simulated not-reached items when the items were answered randomly or left blank, and when the not-reached blanks were scored wrong, scored fractionally correct, and treated as not-administered. Not-reached items were selected randomly and were not related to ability. When not-reached items were left blank and scored wrong, the correlation between the estimated and true item parameters were smaller (and the root mean square errors were larger) when the items were ordered randomly than when they were ordered by increasing difficulty. For example, when 15% of the data was not-reached (the most extreme condition), the true slope parameters had a correlation of .81 with the slope parameters of the items ordered by difficulty but only .46 with the items in random order. The difficulty parameters were generally less affected but showed the same relationship; correlations of .96 when items were ordered by difficulty and .80 when they were in random order. The slope and difficulty parameters for items at the end of the test were estimated to be higher than their true values. The correlations between the true and estimated slopes were higher when the not-reached blanks were scored as fractionally correct, and somewhat higher still when treated as not-administered. The difficulty correlations were not helped by the fractional correct scoring, but were much higher under the not-administered treatment. In interpreting Oshima's results, it is

important to remember the data were missing at random and missing data was unrelated to ability.

This body of research (Lord 1974; Lord, 1980; Ludlow & O'Leary, 1999; Oshima, 1984) suggests not-reached items are generally best treated as not-administered in the item parameter calibration. However, Zimowski et al. (1996) and de Gruitjer (1988) noted in a context of equating two forms where many of the items were literally not-administered, that if the groups answering different sets of items differed in ability, separate population distributions for each group should be used for MML estimation of item parameters. This could be extended to conclude that, if a single population distribution is used in MML, and the examinees who complete the test have a considerably different ability distribution than those who do not reach all items, inaccurate parameter estimations might be obtained for not-reached items, even if they were treated as not-administered. Further, if the group of students who omits the last item has a different ability distribution than the group that omits the last two items, and so on, it would be impractical to estimate a separate latent ability distributions for each group.

To tie these pieces together, two more theoretical pieces can help in understanding when not-administered and not-reached items can be ignored in the estimation process. Rubin (1976) proposed that missing data can be ignored under both direct maximum likelihood and Bayesian estimation, *if* the data are missing at random and the process of missingness is unrelated to the parameter which is estimated. He showed this mathematically. Extending this, Mislevy and Wu (1988) detailed the implications of Rubin's work for estimating ability under several conditions of missing data (alternate forms, targeted testing, adaptive testing, and not-reached items). Some of this work could be applied to the estimation of item parameters, though the focus of Mislevy and Wu was on the estimation of person parameters. In particular, Mislevy and Wu explained

that direct likelihood estimates of ability (such as maximum likelihood with no Bayesian modifications) will generally be accurate when based only on the items the examinee was administered/reached. However, they noted that the prior distribution should take into account the relationship between ability and not-reached items if Bayesian procedures are used to estimate ability; the use of one prior for all examinees, regardless of items not-reached, could produce inaccurate estimates. This is relevant for estimating items because marginal maximum likelihood uses the ability distribution (the procedure involves marginalizing, or averaging over, the ability distribution in estimating item parameters). Generally, it is assumed that the ability distribution does not change for those answering different items. If the ability of examinees presented with (or reaching) item A is quite different than the ability of those presented with (or reaching) item B, it could be problematic to estimate parameters for both items using the same ability distribution. This is not an issue for joint maximum likelihood because only the students who were presented with (or reached) the item are effectively utilized in estimation of that item's parameters.

The present study provides some examples to illustrate these issues. The examples can serve as case studies to alert readers to potential problems; they can not be used to draw conclusions about the frequencies of such cases in actual situations. In fact, for illustration purposes, the association between ability and not-reached items was chosen to be quite high (recall that if there is no relationship there is no theoretical problem), likely more extreme than most real-life data sets.

Method

Using a one-parameter logistic (1PL) item response model, data were simulated for this study. The 1PL model was chosen for data simulation, both for simplicity and because

commercial software is readily available for calibrating 1PL item parameters through both JMLE and MML. Ability levels for 2000 examinees were drawn from a normal distribution with a mean of 0 and a standard deviation of 1. Sixty items were simulated, with difficulty parameters evenly spaced between -1.5 and 1.5. Responses to all 60 items were simulated for each examinee using the 1PL model (*complete data*). Then, responses to some of the last 10 items were changed to not-reached for some examinees using two methods: (1) less-able examinees were less likely to finish the test, with the number of items not-reached at the end of the test correlated -0.84 with ability (*systematically missing*) or (2) items were randomly omitted (*randomly missing*). The proportion of examinees with responses changed to not-reached for each item was kept the same for the two datasets (systematically and randomly missing), and these proportions are shown in Table 1.

Table 1: Proportion of Simulees with Missing Data

	Proportion of Missing Data	
Item	Random Condition	Systematic Condition
51	0.40	0.39
52	0.43	0.41
53	0.44	0.44
54	0.45	0.46
55	0.49	0.49
56	0.52	0.51
57	0.53	0.53
58	0.58	0.57
59	0.61	0.61
60	0.68	0.69

Results

Item parameters for each of the three data sets were estimated by JMLE and MML methods. ConQuest (Wu, Adams, & Wilson, 1998) was used for the MML estimation and BIGSTEPS (Linacre & Wright, 1998) was used for the JMLE. Using the results in Table 2, the JMLE and MML methods can be compared for the complete data set, the randomly missing data set, and the systematically missing (correlated with ability) data set. The complete data set, rather than the true values, was used for the comparisons because it captures some of the random error introduced in simulating the data and re-estimating the item parameters. To highlight the findings for the items with missing data (compared to some of the other items), items 41-60 are displayed in Figures 1 and 2. There is no missing data for items 41-50, and the item difficulty estimates from the three data sets are identical in the JMLE case, and close in the MML case. For items 51-60, where there is some missing data, the estimates vary somewhat. In the JMLE case, there is no clear pattern for which of the data sets has larger difficulty estimates; it appears to be random error introduced by the missing data. Similarly, in the MML case the difficulty estimates based on randomly missing data are sometimes higher and sometimes lower than the estimates based on complete data. For the systematically missing data, however, there appears to be a pattern under MML estimation. For the last eight items, the estimates based on systematically missing data (where students with low scores are more likely to be missing) are consistently lower than the difficulties based on complete data.

Table 2
Data with and without Not-Reached Items: Item Difficulty Estimates

Item	True Item Difficulties	JMLE			MML		
		Complete Data	Randomly missing	Systematically missing	Complete Data	Randomly missing	Systematically missing
1	-1.50	-1.574	-1.573	-1.568	-1.499	-1.546	-1.530
2	-1.45	-1.513	-1.512	-1.508	-1.438	-1.486	-1.470
3	-1.40	-1.315	-1.314	-1.309	-1.244	-1.291	-1.275
4	-1.35	-1.281	-1.280	-1.276	-1.210	-1.258	-1.242
5	-1.30	-1.251	-1.250	-1.245	-1.180	-1.228	-1.212
6	-1.25	-1.185	-1.184	-1.180	-1.116	-1.163	-1.148
7	-1.20	-1.179	-1.178	-1.174	-1.110	-1.158	-1.142
8	-1.14	-1.182	-1.181	-1.177	-1.114	-1.160	-1.145
9	-1.09	-1.167	-1.166	-1.162	-1.101	-1.146	-1.130
10	-1.04	-1.112	-1.111	-1.107	-1.049	-1.092	-1.076
11	-0.99	-1.004	-1.003	-0.999	-0.946	-0.985	-0.970
12	-0.94	-0.999	-0.998	-0.993	-0.943	-0.980	-0.964
13	-0.89	-0.987	-0.986	-0.982	-0.935	-0.969	-0.953
14	-0.84	-0.867	-0.866	-0.861	-0.820	-0.850	-0.835
15	-0.79	-0.859	-0.858	-0.853	-0.816	-0.842	-0.827
16	-0.74	-0.749	-0.748	-0.744	-0.713	-0.735	-0.720
17	-0.69	-0.744	-0.743	-0.739	-0.712	-0.730	-0.714
18	-0.64	-0.611	-0.610	-0.606	-0.586	-0.599	-0.584
19	-0.59	-0.604	-0.603	-0.598	-0.583	-0.592	-0.577
20	-0.53	-0.489	-0.488	-0.484	-0.475	-0.479	-0.464
21	-0.48	-0.527	-0.526	-0.522	-0.516	-0.517	-0.502
22	-0.43	-0.494	-0.493	-0.489	-0.488	-0.484	-0.469
23	-0.38	-0.364	-0.363	-0.359	-0.364	-0.356	-0.341
24	-0.33	-0.316	-0.316	-0.311	-0.322	-0.310	-0.295
25	-0.28	-0.249	-0.249	-0.245	-0.260	-0.244	-0.229
26	-0.23	-0.210	-0.209	-0.205	-0.225	-0.205	-0.191
27	-0.18	-0.210	-0.209	-0.205	-0.229	-0.205	-0.191
28	-0.13	-0.175	-0.175	-0.171	-0.198	-0.171	-0.157
29	-0.08	-0.198	-0.197	-0.193	-0.223	-0.193	-0.179
30	-0.03	-0.001	0.000	0.004	-0.033	0.001	0.014
31	0.02	0.016	0.017	0.021	-0.018	0.017	0.031
32	0.07	0.076	0.076	0.080	0.038	0.076	0.089
33	0.13	0.110	0.111	0.114	0.070	0.109	0.123
34	0.18	0.236	0.237	0.241	0.192	0.234	0.246
35	0.23	0.291	0.292	0.296	0.245	0.288	0.300
36	0.28	0.306	0.307	0.311	0.259	0.302	0.315
37	0.33	0.352	0.353	0.356	0.303	0.347	0.359
38	0.38	0.352	0.353	0.356	0.303	0.347	0.359
39	0.43	0.473	0.475	0.477	0.423	0.467	0.478
40	0.48	0.448	0.449	0.452	0.399	0.442	0.453
41	0.53	0.611	0.612	0.615	0.559	0.602	0.613
42	0.58	0.658	0.660	0.662	0.607	0.648	0.659
43	0.63	0.595	0.596	0.599	0.547	0.586	0.597
44	0.68	0.755	0.756	0.758	0.705	0.743	0.753
45	0.74	0.757	0.759	0.761	0.711	0.746	0.756
46	0.79	0.85	0.852	0.854	0.805	0.837	0.847
47	0.84	0.853	0.855	0.857	0.811	0.840	0.849
48	0.89	0.895	0.897	0.898	0.856	0.881	0.890
49	0.94	0.870	0.872	0.873	0.835	0.856	0.865
50	0.99	0.969	0.971	0.972	0.936	0.953	0.962
51	1.04	1.097	1.077	1.112	1.067	1.061	1.058
52	1.09	1.203	1.165	1.206	1.177	1.145	1.148
53	1.14	1.130	1.178	1.090	1.110	1.153	1.029
54	1.19	1.282	1.308	1.284	1.265	1.281	1.217
55	1.24	1.182	1.290	1.163	1.172	1.266	1.094
56	1.29	1.234	1.207	1.186	1.230	1.184	1.114
57	1.34	1.432	1.461	1.430	1.432	1.433	1.352
58	1.40	1.432	1.439	1.399	1.438	1.415	1.317
59	1.45	1.494	1.318	1.490	1.505	1.295	1.399
60	1.50	1.456	1.447	1.349	1.468	1.420	1.245

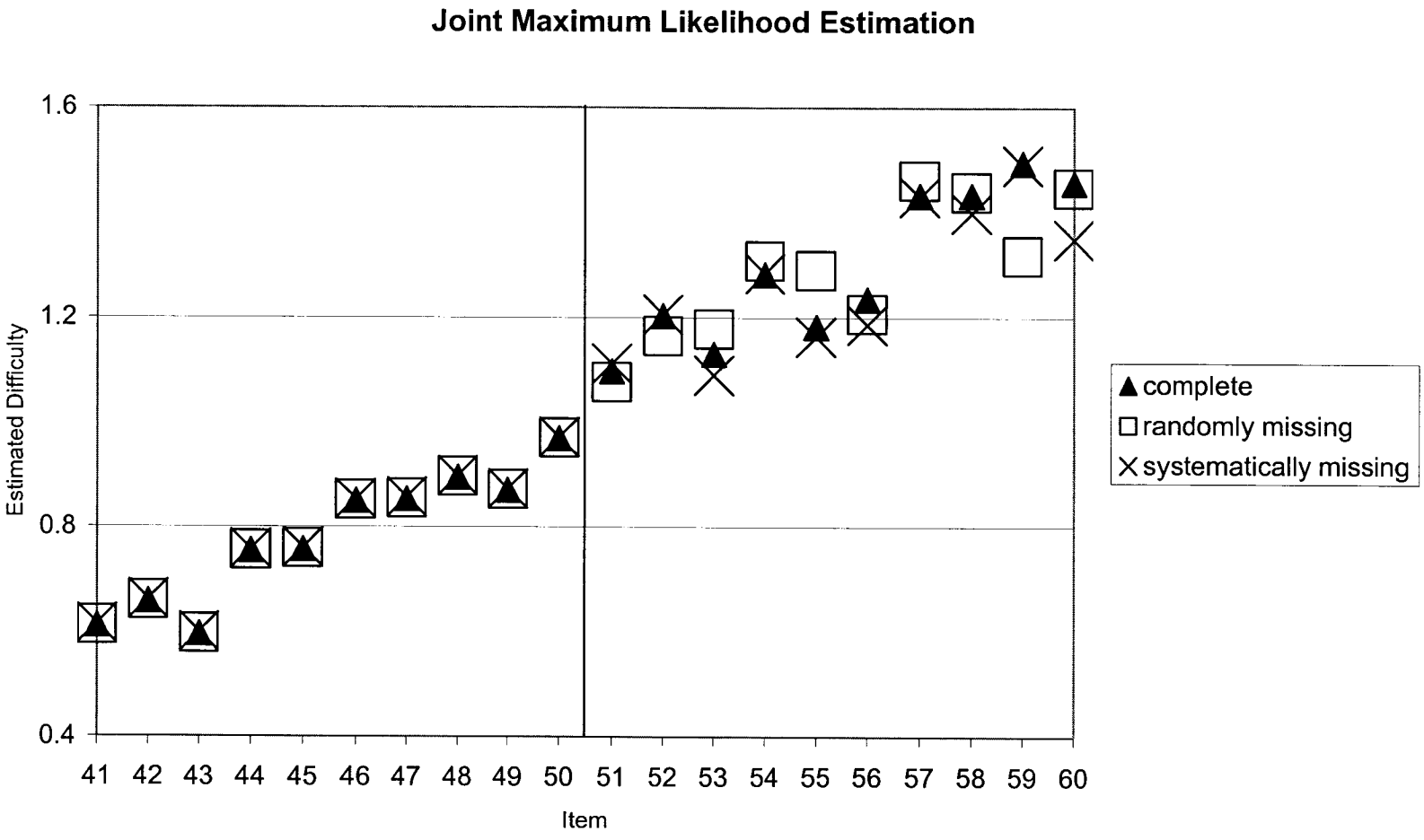

Figure 1. JMLE estimates of item parameters for selected items with no missing data, randomly missing data, and systematically missing data.

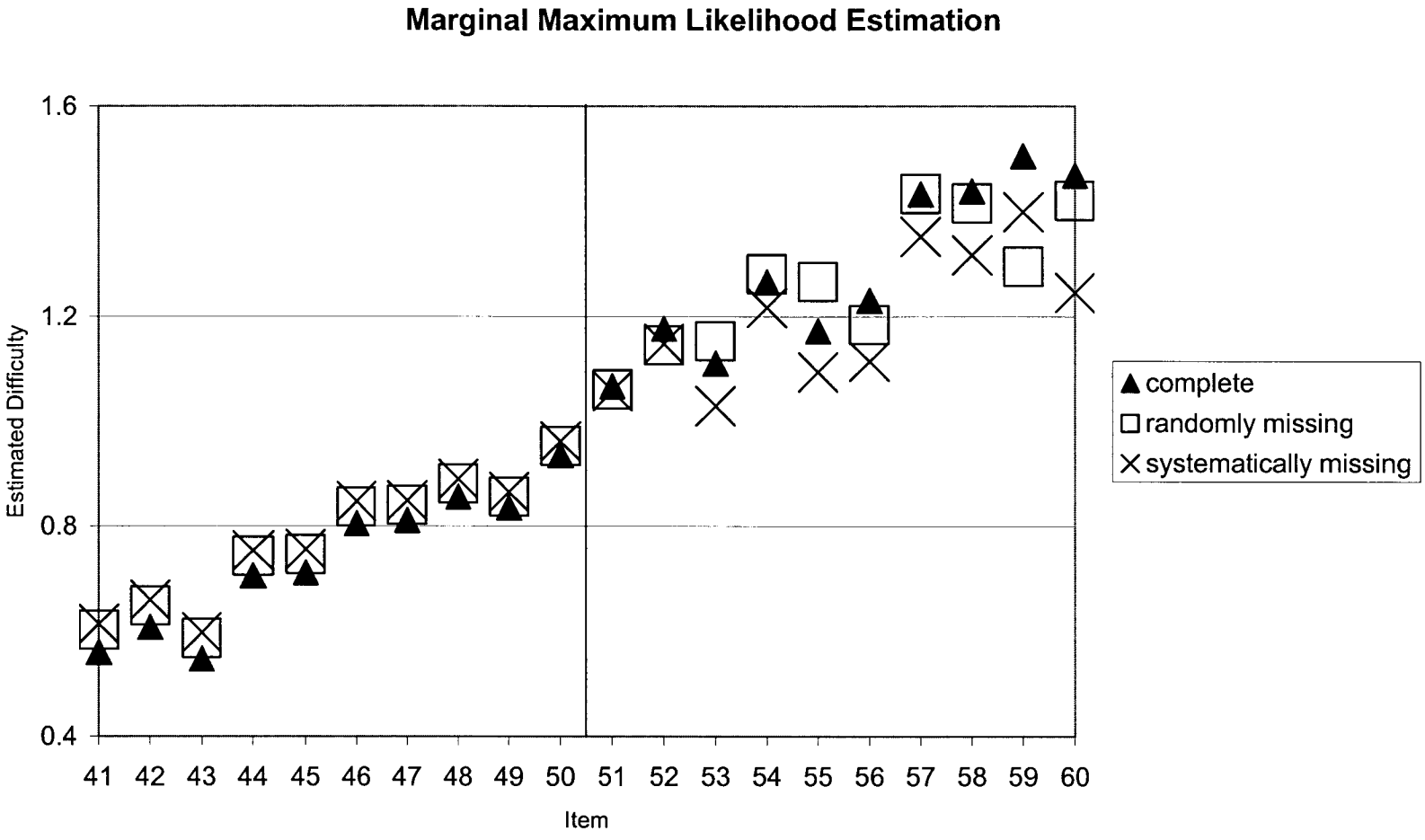

Figure 2. MML estimates of item parameters for selected items with no missing data, randomly missing data, and systematically missing data.

Conclusions

These examples have illustrated that, although JMLE theory allows for ignoring missing data, MML theory allows for ignoring missing data only if it is randomly missing (or if there are known groups for which separate latent distributions can be estimated, as in vertical equating). These examples are only case studies; they do not provide standard errors or even average effects (simulations could be used for that purpose, if desired). More importantly, these examples do not estimate the frequency with which such results are likely to occur under practical conditions. The not-reached-items example is contrived to show a worst case scenario, with high rates of not-reached items and a high correlation between ability and reaching items at the end of the test. If there is little correlation between ability and finishing a test, the problem illustrated here would not be an issue. The purpose of these examples was to display how ignoring missing data could potentially be problematic in item difficulty estimation.

These results also have implications for re-estimating item parameters for tests administered adaptively. Item parameters might be estimated, for example, to check for parameter drift. In adaptive testing, each examinee receives a different mix of test items, and after the initial items the items should be matched to ability. MML estimation with a single latent distribution might be inappropriate in this case, because the ability distribution of those who were administered a particular item would differ from the ability distribution of those who were administered a different item. Further research would be needed to see if this presents a meaningful problem in actual usage.

References

De 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, 213-234.

de Gruitjer, D. N. M. (1988). Standard errors of item parameter estimates in incomplete designs. Applied Psychological Measurement, 12, 109-116.

DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML. Applied Measurement in Education, 15, 15-31.

Hambleton, R. K. & Swawinathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer.

Hambleton, R. K., Swawinathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: SAGE.

Linacre, J. M., & Wright, B. D. (1998). BIGSTEPS 2.82 [Computer software]. Chicago: MESA Press.

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. Mahwah, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. Psychometrika, 48, 477-482.

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: Educational Testing Service. [ERIC Document Reproduction Service No. ED 395 017]

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. Journal of Educational Measurement, 31, 200-219.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63, 581-592.

Wood, R.L., Wingersky, M.S. & Lord, F.M. (1976). LOGIST: *A Computer Program for Estimating Examinee Ability and Item Characteristic Curve Parameters.* Princeton, NJ: Educational Testing Service.

Wu, M. L., Adams, R. J., Wilson, M. R. (1998). ACER ConQuest [Computer software]. Melbourne, Australia: Australian Council for Educational Research.

Zimowski, M.. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG [Computer software]. Chicago: Scientific Software International.