Psychological Test and Assessment Modeling, Volume 55, 2013 (4), 335-360

Designing small-scale tests: A simulation study of parameter recovery with the 1-PL

Dubravka Svetina[1], *Aron V. Crawford*[2], *Roy Levy*[2], *Samuel B. Green*[2], *Lietta Scott*[2], *Marilyn Thompson*[2], *Joanna S. Gorin*[2], *Derek Fay*[2] *& Katie L. Kunze*[2]

Abstract

This simulation study investigated the recovery of item and person parameters of the one-parameter logistic model for short tests administered to small samples. A potential problem with such small scale testing is the mismatch between item and person location parameter distributions. In our study, we manipulated the match of these distributions as well as test length, sample size, and item discrimination. Results showed the degree of mismatch likely to occur in practice has a relatively modest effect on parameter recovery. As expected, accuracy in parameter estimation decreased as sample size and test length decreased. Nevertheless, researchers investigating small scale tests are likely to view parameter recovery as acceptable if a study has at least 100 subjects and 8 items.

Key words: parameter recovery, mismatch of item and person location distributions, small-scale assessments, item response theory, simulation study

[1] *Correspondence concerning this article should be addressed to:* Dubravka Svetina, PhD, 201 N. Rose Avenue, Indiana University, Bloomington, IN 47405; email: dsvetina@indiana.edu.

[2] Arizona State University

Psychometric methods based on item response theory (IRT) have been extensively investigated for large-scale assessments characterized by administering long tests to large samples. These applications frequently concern high-stakes testing, which necessitates a high level of quality of the parameter estimates. Often, in such situations, the person parameter and item location distributions are assumed to be comparable or matched in the sense that the distribution of locations of the test items matches the distribution of person parameters in the population of examinees; typically both distributions are assumed to be standard normal. Considerably less research has been conducted investigating IRT methods where a mismatch between item and person parameter distributions exists. Such situations are more likely to be found and be of concern in small-scale scenarios characterized by a limited number of items (e.g., 8 to 20 items) and small samples (e.g., between 100 and 200 subjects). Examples of low-stakes contexts in which shorter scales are commonly administered to small samples include the development of tests for research purposes (e.g., piloting of tests, general information of item quality), psychometric investigations of non-cognitive scales, and tests used as screening tools. Standard principles of estimation would suggest that the quality of the parameter estimation in such situations would suffer relative to the large-scale applications. However, in small-scale applications, researchers may be willing to accept estimates that are less accurate compared to those in large-scale testing applications (Hambleton, 1989; Linacre, 1994).

Intuitively, with a one-parameter logistic model, we would expect the quality of parameter estimates to degrade to the extent that item and person location parameter distributions failed to match because an individual item is maximally informative for person parameter estimation and an individual person is maximally informative for item parameter estimation when the person and item are located at the same place on the latent continuum (i.e., when $\theta_i = b_j$, in the notation introduced below). The current study focuses on examining the impact of item-person mismatch on parameter estimation for small-scale assessments utilizing information functions in conjunction with simulation techniques.

In the introduction, we discuss the choice of item response models for small-scale tests and why item-person mismatches are likely to be most problematic with this type of test. We then review the relevant literature on IRT estimation, particularly as it applies to short-scale tests. Finally we describe the objective of the study.

Choice of IRT model for small-scale tests

The majority of dichotomous IRT applications apply one of three models, which vary in complexity in terms of restrictions on parameters. The three-parameter logistic (3-PL) model specifies the probability of person *i* endorsing an item *j* ($X_{ij} = 1$) as:

$$P\left(X_{ij}=1|\theta_i,a_j,b_j,c_j\right)=c_j+\left(1-c_j\right)\frac{exp^{a_j\left(\theta_i-b_j\right)}}{1+exp^{a_j\left(\theta_i-b_j\right)}},$$

where θ_i is the person parameter for person *i*, b_j is the location parameter for item *j*, a_j is the discrimination parameter for item *j*, and c_j is a lower-asymptote parameter for item *j*. When c_j is fixed to 0 for all items, the 3-PL simplifies to a two-parameter logistic (2-PL) model. When all of the discrimination parameters are further constrained to be equal across items, the model simplifies to a one-parameter (1-PL) model, with its special case being that of the Rasch model (Fischer, 1974; Wright & Stone, 1979). In the 1-PL model, the discriminations are frequently assumed to equal 1 and the metric (variance) of the person parameters is estimated. An alternative is to estimate a single value for the discrimination value assuming the metric of the person parameters is known.

Model selection and parameter estimation decisions are often affected by the intended context and purpose of testing. For example, in low-stakes testing situations, simple summed scores are likely to be preferred in that they require no specialized software or training on the part of the examiner. Further, simpler models that generally carry less stringent sample size requirements are also likely to be favored. Given these considerations, we focus our investigation on the 1-PL model.

Type of test and item-person mismatch

Researchers who develop small-scale tests for low-stakes applications are likely to encounter mismatches between the locations of item and person parameters on the latent continuum. This item-person mismatch is more likely to occur when a researcher with limited resources uses a convenience sample. For example, a researcher creates a scale with true-false items that are intended to measure moderate to severe depression. The researcher has access to a convenience sample, which includes depressed individuals, but few, if any, severely depressed persons. As a result, the sample distribution is likely to be lower on the depression continuum relative to the items that were designed to make distinctions among levels of depression for a truly depressed population.

The effect of item-person mismatch on parameter estimation is of less interest with high-stakes tests, which typically are administered in the scale development stage to samples that are similar to their intended population. On the surface, an exception appears to be for adaptive testing situations in which the item pool characteristics may not provide optimal information for special populations of examinees (Chen, Hou, Fitzpatrick, & Dodd, 1997; Dodd, Koch, & De Ayala, 1993; Gorin, Dodd, Fitzpatrick, & Sheih, 2005). The small number of items that typically are administered in adaptive tests makes applications with these tests more akin to those for small-scale tests. However, because the initial calibration of item parameters for high-stakes adaptive tests generally involves large samples of examinees, problems with item-person mismatch are probably minimized. The current study focuses on examining the impact of item-person mismatch on parameter estimation in small-scale IRT contexts, utilizing information functions in conjunction with simulation techniques.

Item estimation for small-scale assessment

A number of factors may affect the precision of the item and person parameter estimates, including the researcher's choice among IRT models, estimation methods, software programs, and shapes of item and person parameter distributions (Hambleton, 1989). The literature contains limited recommendations about minima for sample size and test length, and these recommendations are spread among a variety of IRT models and estimation algorithms under a range of conditions. Accuracy of estimation is tied generally to the number of estimated parameters, and consequently sample-size and test-length recommendations generally increase as one moves from less complex to more complex IRT models.

Previous research has shown that based on analytical results utilizing the Rasch model, stable estimates may be obtained with 30 well-targeted examinees and 30 well-fitting items (Linacre, 1994; Wright & Douglas, 1975, 1976; Wright & Stone, 1979). Linacre (1994) defined boundaries of stable item calibrations as either .5 or 1 logit away from the point estimate, as the author's purpose was to stay within a grade level for any one set of items. Linacre (1994) and others remind us that these requirements could be lowered even further in situations where researchers are willing to accept a relaxed level of error given the purpose for the test.

Early literature suggested that sample sizes as small as 100 might be sufficient to achieve satisfactory calibrations when the items on the test are positioned within a logit or two of the mean ability of the group (Wright, 1977). Similarly, Lord (1983) suggested that the Rasch model can be justifiably used in estimating person parameters in shorter tests (10 and 15 items) when the available sample size is less than 100 or 200.

The quality of estimation and associated recommendations are also necessarily tied to the estimation method (Hambleton, 1989). More recently, Wang and Chen (2005) investigated item parameter estimation using joint maximum likelihood (JML) estimation as implemented in the WINSTEPS program (Linacre, 2010). The test length ranged from 10 to 60 items and the sample size varied from 100 to 2000. Wang and Chang (2005) found that WINSTEPS underestimated the location of items in the lower tail of distribution and overestimated the locations of items in the upper tail of distribution. The maximum bias found was in tests with 10 items; bias in longer tests was found to be negligible (bias for item parameters ranged from .272 to .074 for tests 10 and 60 items, respectively). Additionally, Xiong, Lewis, and Mingmei (2009) found that marginal maximum likelihood (MML) provided better accuracy than conditional maximum likelihood (CML) for estimating 1-PL model with as few as 5 items and 50 examinees. Hulin, Lissak, and Drasgow (1982) conducted simulations using 2-PL and 3-PL models with 15, 30, and 60 items and 200, 500, 1,000, and 2,000 simulated examinees using JML. They concluded that accurate recovery of item characteristic curves, in which the average root mean squared error (RMSE) was less than .05, can be achieved with 30 items and a sample size of 500 for a 2-PL model, and 60 items and a sample size of 1,000 for a 3-PL model.

Drasgow (1989) found that for a given sample size and number of items in the 2-PL model, MML estimation was superior to JML. Harwell and Janosky (1991) investigated tests with 15 or 25 items administered to 75, 100, 150, 250, 500, or 1,000 simulees using MML. They found with the 2-PL model that accurate item parameter estimates (RMSEs < .20) were obtained with at least 15 items and 250 examinees. Interestingly, they reported RMSE values of .21 for item location parameter recovery when there were 15 items and 100 examinees, but did not recommend using samples that small.

A similar degree of error was described by Stone (1992) as "generally precise and stable" (p. 1). Stone used MML to fit the 2-PL model with 10, 20, or 40 items and 250, 500, and 1,000 simulated examinees. He found acceptably accurate estimates of item location parameters (RMSE < .22) with as few as 10 items and 250 examinees. Drasgow (1989) found that as few as 5 items and 200 persons were required for parameter estimates with reasonably small standard errors (SEs < .30) for some attitude scale applications using the 2-PL. The estimation of discrimination parameters tends to be more problematic than estimation of item location parameters (Drasgow, 1989; Harwell & Janosky, 1991; Hulin, et al., 1982; Stone 1992). Accordingly, item and sample requirements for 1-PL models may be even lower than those reported in the previously discussed studies based on 2-PL models.

As compared to suggested guidelines for accuracy of item parameter estimates, researchers seem less stringent about the acceptable level of error for person parameter estimates. Hulin et al. (1982) described an RMSE of .38 for person parameter recovery as "very precise," and Stone (1992) described person parameter RMSEs around .40 as "small." Researchers oriented toward person evaluation would potentially set different standards of error acceptability. As suggested earlier, studies examining the precision of item and person parameter recovery are often tied to the choice of the IRT models under consideration as well as the context within which they are situated (e.g., with the purpose of comparing software performance in parameter recovery, equating, etc.). Related psychometric work has been conducted with the purpose of parameter recovery in polytomous IRT models (e.g., Ankenmann & Stone, 1992; Childs & Chen, 1999; Reise & Yu, 1990; Wollack, Bolt, Cohen, & Lee, 2002), mixed item format examinations (e.g., Jurich & Goodman, 2009), software comparisons (e.g., Childs & Chen, 1997; Jurich & Goodman, 2009; Mislevy & Stocking, 1989; Yen, 1987) and equating (e.g., Baldwin, Baldwin, & Nering, 2007; Kim & Cohen, 2002). However, most of these studies used larger sample sizes and most have not focused on the issues related to the mismatch between the distributions of item difficulty and person parameters in a systematic way.

Objective of the study

A popular perception among psychometricians is that sample size and test length requirements for IRT estimation have been intensively researched; nevertheless, we found only a small number of studies that focused on the lower limits of sample size and test length. In addition, reaching conclusions based on these studies is problematic in that they tend to use different models, estimators, and/or different combinations of items and sample sizes. Furthermore, none of these studies examined the effects of mismatch be-

tween person and item parameter distributions. Our work is primarily concerned with the estimation accuracy of both person and item location parameters when the underlying distributions of the two are not aligned, a situation that in practice might be found in low-stakes scales. Additional research is needed to offer guidelines for practitioners who are developing low-stakes scales with limited resources. The current study expands on previous literature by focusing on recovery of model parameters with short scales and small samples in which person parameters fail to match the item parameters.

Method

In this section, we describe the conditions initially investigated (referred to hereafter as *original conditions*), including the data generation process, the manipulated factors underlying these conditions, and the criteria with which we evaluated the accuracy and efficiency of parameter recovery. Based on the findings from this investigation, we explored additional conditions that are described following the presentation of the initial findings in the Results section.

Data generation

Dichotomous item responses were generated based on the 1-PL model using code written by the authors in R (R Core Development, 2006). For each condition, appropriate item and person parameters were used (see section on *Manipulation of Factors* for detail on item and person values) to simulate the item responses using the *sim* function in the `irtoys` package in R (Partchev, 2010). Once the datasets of 0s and 1s were simulated, they were analyzed to obtain item parameter estimates using MML as implemented in BILOG (Mislevy & Bock, 1982). We used the `irtoys` package in R again (Partchev, 2010) to call the BILOG program to fit the model (using default settings) and to obtain item and person parameter estimates (i.e., expected a posteriori (EAP) estimates of person parameters). One thousand data sets were generated and analyzed for each condition. The results for the 1000 replications were then summarized within and across conditions.

Manipulation of factors

We manipulated four factors using a crossed design which produced a total of 16 original conditions.

Match of person parameter (θ) and item locations (*b*). The person parameters (θ) were generated from a population assuming a normal distribution with a mean of 0 and a variance of 1 for all original conditions. Item location parameters (*b*) also were generated from normally distributed populations. For matched conditions, the means and variances of the normally distributed location parameters were 0 and 1. For the mismatched conditions, the means and variances were .5 and 1. Additional conditions examining more extreme levels of mismatch were examined and are discussed in a subsequent section. In

the generation of item location parameters, the normal distribution was divided into 8 equal probability bins for the 8-item conditions and 16 probability bins for the 16-item conditions. That is, any one item had equal probability of being placed in any one bin, given the bin boundaries. Bin boundaries in the 16-item conditions were -3, -1.53, -1.15, -0.89, -0.67, -0.49, -0.32, -0.16, 0, 0.16, 0.32, 0.49, 0.67, 0.89, 1.15, 1.53, and 3. In the 8-item conditions, every other boundary was used (i.e., -3, -1.15, ... , 3). Item locations were generated for any one replication such that one item was included in every bin, resulting in a distribution of *b* values for each replication that approximated a standard normal distribution. An identical approach was used to generate scales with shifted location parameters, except .5 was added to generated locations.

Number of scale items (*J*). The purpose of the study was to investigate sample size requirements for scales with relatively small numbers of items. Accordingly, we explored scales with 8 or 16 items.

Sample size (*N*). Past research suggests that a sample size of 100 can be adequate with 1-PL model. Accordingly, we chose a sample size of 100 to represent a small sample and 500 for a large sample.

Discrimination (*a*). More subjects are required to the extent that items are less discriminating. Thus, it was important to manipulate discrimination. For any scale, all items on a scale had a discrimination of either .7 or 1.3.

Criteria for evaluation of parameter recovery

We computed three indices to assess accuracy of item and person parameter estimation across replications: bias, root mean squared error (RMSE), and correlations. These indices were computed both within bins and across all bins. Bias was assessed by computing the mean difference between the estimated and true values of the parameters. The bias for the *b* parameters from within a bin, across replications is given as

$$Bias_{bin} = \sum_{rep=1}^{1000} \frac{\left(\hat{b}_{rep} - b_{rep}\right)}{J},$$

where $\hat{b}_{rep}$ and b_{rep} are the estimated and true *b* values for the item for the replication. RMSEs were calculated by taking the square root of the mean of squared deviations of estimated parameter values about their true values. For example, the RMSE for the location parameters for a particular bin was calculated as

$$RMSE_{bin} = \left[\sum_{rep=1}^{1000} \frac{\left(\hat{b}_{rep} - b_{rep}\right)^2}{1000} \right]^{\frac{1}{2}}.$$

Correlations were computed between the estimated and the true parameters. In our results, we focus on the bias and RMSEs in that the correlations were not particularly sensitive to differences in manipulated factors.

Results

We first present results for the recovery of person parameters, followed by results for the recovery of item location parameters that include match/mismatch conditions. Due to space limitations, we present selected results in tabular or graphical form, but include all conditions in our syntheses of results. Complete tables of results are available upon request from the authors.

Person parameter recovery

For each condition, the RMSEs and bias were computed within each bin. Broadly speaking, for mismatched conditions in which item locations were uniformly shifted upward, the person parameter recovery was slightly worse than in corresponding matched conditions for lower values of θ and slightly better for higher values of θ (difference in RMSEs was never larger than .01).

To summarize the recovery of person parameter estimates across bins, two tables are provided. In panel (a) of Table 1, person parameter recovery for two selected conditions (one mismatched and one matched) for all bins is shown. The reported conditions had item discriminations of .7, 8 items, and sample size of 100, and were representative of the general pattern of findings.

Panel (a) in Table 1 reveals that only slight differences in the performance were noted between mismatched and matched conditions. Neither of the conditions consistently outperformed the other with the largest differences found in the extreme tails of the distribution. Mismatched condition performed better in the upper (positive) extreme of the distribution than the matched condition with difference in RMSE by .02 and bias of .03. The opposite pattern and magnitude of performance were found in the lower (negative) extreme of the distribution, where the matched condition yielded lower RMSE and bias than mismatched condition by .02 and .03, respectively. Generally, for both mismatched and matched conditions, the poorest parameter recovery occurred in the extreme ends of the distribution as evidenced by both RMSE and bias, whereas the most accurate recovery occurred in the center of the person distribution.

Panel (a) of Table 2 summarizes the comparative patterns of results in the match and mismatch across all conditions. In this table, the number of reported bins was reduced to allow the inclusion of the other manipulated factors. Inspection of the table reveals slight differences between matched and mismatched conditions (holding other design factors constant) across the values of the other design factors.

Table 1:
RMSEs and Bias for Selected Conditions for All Bins.

		Panel (a): *Person Parameter Estimates*							
Condition		(-∞, -1.15)	(, -.67)	(, -.32)	(, 0)	(, .32)	(, .67)	(, 1.15)	(1.15, ∞)
Matched, 8 items, $N = 100$, Discrimination = .7	RMSE	1.08	0.69	0.57	0.52	0.51	0.57	0.70	1.08
	Bias	0.93	0.48	0.26	0.09	-0.09	-0.27	-0.48	-0.94
Mismatched, 8 items, $N = 100$, Discrimination = .7	RMSE	1.10	0.68	0.56	0.51	0.52	0.58	0.70	1.06
	Bias	0.96	0.48	0.25	0.08	-0.11	-0.28	-0.48	-0.91
		Panel (b): *Item Location Parameter Estimates*							
Condition		(-3, -1.15)	(, -.67)	(, -.32)	(, 0)	(, .32)	(, .67)	(, 1.15)	(1.15, 3)
Matched, 8 items, $N = 100$, Discrimination = .7	RMSE	0.58	0.46	0.40	0.35	0.34	0.39	0.42	0.59
	Bias	-0.12	-0.07	-0.04	-0.02	0.01	0.01	0.05	0.13
Mismatched, 8 items, $N = 100$, Discrimination = .7	RMSE	0.53	0.37	0.36	0.35	0.42	0.47	0.61	0.74
	Bias	-0.08	-0.02	-0.01	-0.02	0.06	0.07	0.12	0.15

Note: For mismatched conditions in this table bins are shifted up by .5. Bin boundaries are indicated for the most outer bins. For space constraints, only the upper boundaries for bins two through seven are listed.

Table 2:
RMSEs for Person and Item Location Parameter Estimates for the 16 Original Conditions.

		Panel (a): *RMSEs for Person Parameter Estimates*							
		$a = .7$				$a = 1.3$			
		(-∞, -1.15)	(-.32, .32)	(1.15, ∞)	M_G	(-∞, -1.15)	(-.32, .32)	(1.15, ∞)	M_G
	N								
Match	100	1.08	0.52	1.08	**0.72**	0.74	0.49	0.73	**0.56**
	500	1.07	0.52	1.07	**0.71**	0.74	0.49	0.73	**0.56**
Mismatch	100	1.10	0.52	1.06	**0.71**	0.81	0.50	0.70	**0.57**
	500	1.09	0.52	1.05	**0.72**	0.80	0.49	0.69	**0.57**
Match	100	0.83	0.50	0.83	**0.61**	0.55	0.40	0.55	**0.45**
	500	0.82	0.50	0.82	**0.61**	0.54	0.40	0.54	**0.44**
Mismatch	100	0.85	0.51	0.80	**0.61**	0.62	0.41	0.52	**0.46**
	500	0.85	0.50	0.80	**0.61**	0.60	0.40	0.50	**0.45**

		Panel (b): *RMSEs for Item Location Parameter Estimates*							
		$a = .7$				$a = 1.3$			
		(-3, -1.15)	(-.32, .32)	(1.15, 3)	M_G	(-3, -1.15)	(-.32, .32)	(1.15, 3)	M_G
	N								
Match	100	0.58	0.35	0.59	**0.44**	0.33	0.21	0.35	**0.26**
	500	0.21	0.15	0.22	**0.17**	0.14	0.09	0.14	**0.12**
Mismatch	100	0.53	0.38	0.74	**0.48**	0.27	0.22	0.42	**0.26**
	500	0.17	0.15	0.25	**0.17**	0.12	0.10	0.18	**0.11**
Match	100	0.43	0.33	0.44	**0.38**	0.33	0.21	0.31	**0.25**
	500	0.18	0.15	0.19	**0.16**	0.14	0.09	0.14	**0.11**
Mismatch	100	0.38	0.35	0.52	**0.37**	0.26	0.22	0.38	**0.24**
	500	0.17	0.15	0.21	**0.16**	0.11	0.10	0.16	**0.11**

Note. M_G is the grand mean (average across all bins). For mismatched conditions in this table bins are shifted up by .5.

The largest differences in RMSEs were found in conditions with a sample size of 100 and discrimination of 1.3, where matched conditions outperformed the mismatched conditions at the lower end of distribution for 8 and 16 items by .07 and .07, respectively. The largest discrepancies for mismatched conditions outperforming matched conditions were found in the upper end of distribution in conditions with a sample size of 500 and discrimination of 1.3 for 8 and 16 items (difference of .04 and .04, respectively). In addition, means across all bins are reported. We can see that the marginal effect of mismatch was negligible because gains in recovery at one end of the distribution were washed out by losses in recovery at the other end of the distribution; on average, matched conditions performed equally well or better than mismatched conditions (difference in means for RMSEs was never larger than .01).

Considering the effects of the other manipulated factors, the lowest RMSE values were obtained in conditions having 16 items and discrimination parameters of 1.3, followed by conditions with the same discrimination but only 8 items. Within each discrimination level, increasing the number of items improved the recovery of the person parameter estimates. Similarly, when the number of items was held constant, increasing the discrimination parameter improved the recovery of the person parameter estimates. In contrast, sample size had a negligible impact on the recovery of θ estimates.

In summary, this mismatch between person parameter and item location distributions had relatively weak effects on θ recovery, the number of items and item discrimination had fairly strong effects on θ recovery, and the effects of sample size appeared relatively weak. Consistent with principles of expected a posteriori estimators (Bock & Mislevy, 1982), person parameter estimation was the most accurate in the center of the distribution. In the lower tail of the θ distribution, estimates were positively biased. Conversely, in the upper tail, estimates were negatively biased. The magnitudes of these biases were comparable; due to the symmetrical nature of the distributions, the marginal biases across bins were equal to or less than .01 for matched conditions and were less than .14 for mismatched conditions.

Item location parameter recovery

RMSEs. Generally, the mismatch between the item parameter distribution and the person parameter distribution impacted the recovery of item location parameters more than person parameter estimates. As illustrated for one condition in panel (b) of Table 1, the recovery of item location estimates in the lower tail of the item parameter distribution was better in mismatched conditions than in matched conditions. Here, the difference in RMSEs ranged from .04 to .09, with the largest difference found in the area of -1.15 to -0.67 on the latent scale. However, in the upper tail of the distribution, the recovery of item location estimates was better in matched conditions than in mismatched conditions. The difference in RMSEs ranged from .08 to .19, with the largest difference found in the area of 1.15 to 0.67 on the latent scale. In other words, the estimates of item location parameters were better if a greater proportion of simulees had θ values that were in the same region of the distribution as the location parameters. When averaging across all

bins, the matched conditions were approximately equivalent to their mismatched counterparts.

Bias. Overall, negligible bias was observed across all conditions, especially at the center of the item distribution where bias was essentially zero. When bias was found at the lower levels of the item parameter distribution, location estimates were negatively biased, whereas at the upper levels of the item parameter distribution they were positively biased (see panel (b) of Table 1). This was true for matched and mismatched conditions with varying sample sizes. Among the matched conditions, the largest bias was found in the condition with .7 discrimination, sample size of 100, and 8 items; bias in this condition ranged from -.12 to .13. Overall, the largest amount of bias was found in the condition with .5 mismatch, .7 discrimination, sample size of 100, and 8 items, where bias ranged from -.08 to .15.

Conditions to examine degree of mismatch of item and person parameters

Based on results for original matched versus the mismatched conditions, we developed a *local sample size* hypothesis. Specifically, we speculated that the quality of item parameter estimation was directly related to the relative number of subjects with θ values near the item's *b* parameter. To assess the local sample size hypothesis, we investigated a number of additional conditions. The first set of additional conditions increased the mismatch between the person and item parameter distributions. This was achieved by generating item parameters from a normal distribution with mean of 1.0 rather than .5 as it was the case in the previous mismatch conditions or 0 in match conditions.

The results with no mismatch, a mismatch of .5, and a mismatch of 1.0 are presented in the first, second, and third panels of Figure 1, respectively. Relatively small differences were observed between matched and mismatched conditions at the center of the person distribution. Furthermore, for conditions with a sample size of 500, differences between matched and mismatched cases were fairly small, even in the tails of the distribution. Mismatch had the greatest effect when sample size was 100 and in the tails of the distribution. In other words, as the distribution of item parameters was shifted increasingly to the right on the person parameter distribution, RMSEs for item parameters increased most dramatically in the bins with relatively fewer people. These findings were in support of the local sample size hypothesis.

Conditions with mixture distributions to assess the local sample size hypothesis

We further investigated the local sample size hypothesis by including conditions in which the underlying distribution of person parameters was mixed normal. Data were generated for simulees based on mixed normal distributions with two subpopulations – one subpopulation with a negative mean and the other with a positive mean. If the local sample size hypothesis was correct, the concave shape of the RMSE plots (with item

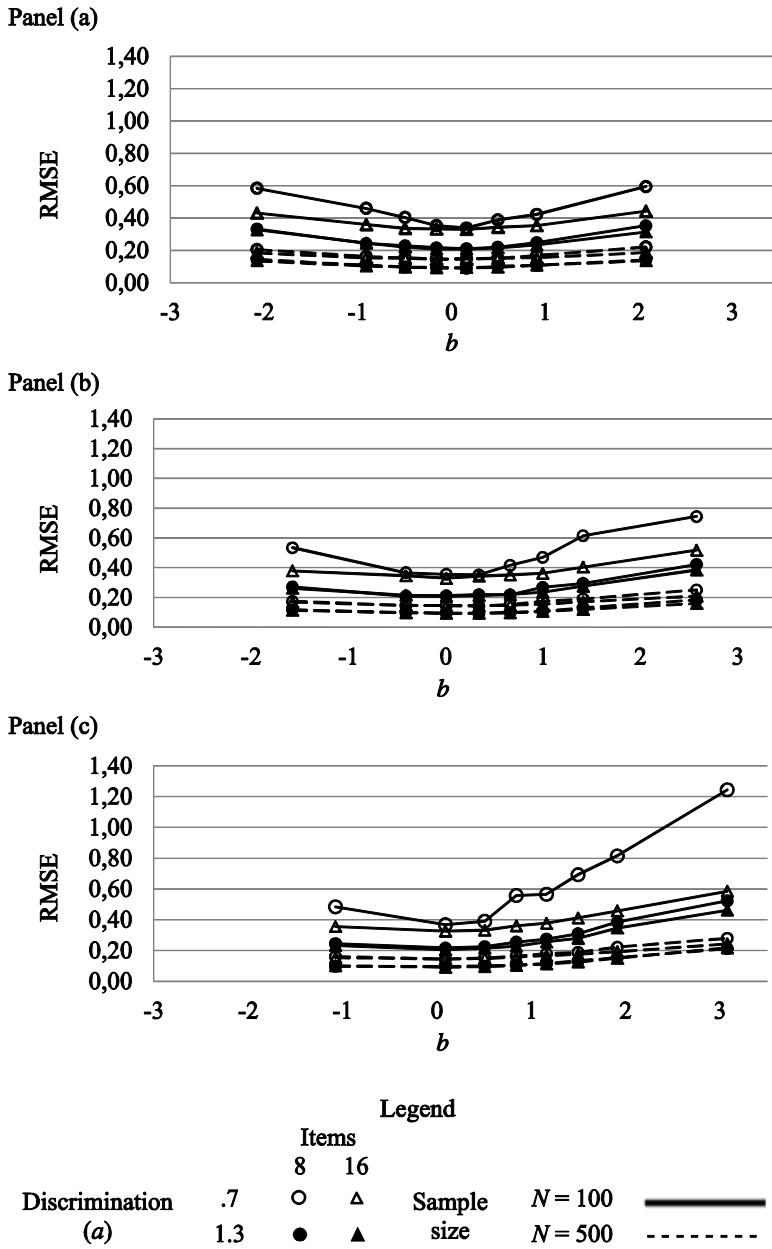

Figure 1:
RMSEs for item location parameters for the matched, mismatched by .5, and mismatched by 1.0 conditions, respectively. Markers are plotted at bin midpoints.

locations on the abscissa) that was previously found would be inverted to form a convex shape. In other words, the lowest RMSEs would be in the tails where relatively more simulees were located, and the highest RMSEs would be in the middle where very few simulees were located.

We manipulated the degree to which the two subpopulations differed across conditions. Manipulated variables included the means for the two subpopulations (-.5 and .5, -1 and 1, -1.5 and 1.5, or -2 and 2), the common variance for the two subpopulations (.0001 to 1), and the proportion of simulees generated from each distribution (.5 and .5, or .2 and .8). We explored a large number of mixture conditions, but present only three of them to illustrate our findings.

Uneven mixture condition. Figure 2 shows the results from a condition where the underlying θ distribution was a mixed normal distribution in which 20% of the simulees came from a *N*(-.5, .0001) distribution located below most of the items, and 80% came from a *N*(.5, .0001) distribution located above most of the items. Each replication of this condition had a sample size of 500, 8 items, and item discriminations of 1.3. Given the small variance used here, the simulees from the two subpopulations of the mixture distribution were almost exclusively along the latent distribution that corresponded to θ

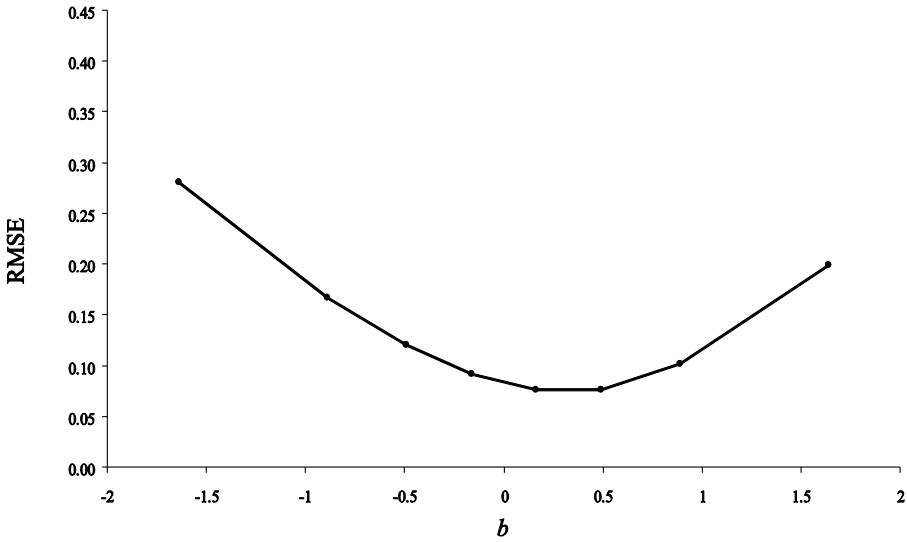

Figure 2:
Item location parameter RMSEs for a mixture distribution condition with sample size of 500, 8 items, and discrimination of 1.3. The underlying θ distribution was mixed normal such that 20% of the simulees were drawn from ~*N*(-.5, .0001) distribution and 80% were drawn from ~*N*(.5, .0001) distribution. Markers are plotted at the mean of the true *b* parameters within each bin.

ranges from -0.67 to -0.32 (bin 3) and from 0.32 to 0.67 (bin 6).[3] The asymmetrical shape of Figure 2 provided some support for the local sample size hypothesis, but further inspection suggested the hypothesis was not fully supported. The items located between 0.32 and 0.67 on the latent continuum (in bin 6) where 80% of the simulees were located had small RMSEs, but were approximately equivalent to the RMSEs in the adjacent bin 5 (0 to 0.32) where few if any simulees were located. Overall, and counter to our speculation based on the local sample size hypothesis, the concave shape seen in previous graphs was preserved in this condition, except for the noted asymmetry.

Extreme mixture condition. Figure 3 shows results from a more extreme mixture distribution, where 50% of the simulees came from a $N(-2, .01)$ distribution and 50% came from a $N(2, .01)$ distribution. This condition included 8 items, a sample size of 500, and item discriminations of 1.3. Given the small standard deviation, the simulees θs were almost exclusively in bin 1 (range $-\infty$ to -1.15) for one of the subpopulations or in bin 8 (range 1.15 to ∞) for the other subpopulation.

If the local sample size hypothesis was correct, the items located in the middle of the latent continuum would be poorly estimated for this condition, whereas those located in the tails would be better estimated. Surprisingly, the item parameters in the middle were still recovered more accurately than those in the tails based on the RMSEs and as shown in Figure 3. Also, as shown in Figure 3, the degree of bias increased substantially as the true *b*s differed from 0.

Given these surprising results, we futher evaluated this condition by assessing the efficiency (i.e., stability) of the item location estimates. The results for the RMSEs, which assess accuracy, could have been primarily a result of bias rather than efficiency. Accordingly, we chose to assess efficiency independently. Efficiency was computed (for each bin) as:

$$Efficiency_{bin} = \frac{\sum_{rep=1}^{1000}\left[\hat{b} - \frac{\sum_{rep=1}^{1000}\left(\hat{b}-b\right)}{1000}\right]^2}{1000}.$$

As shown in Figure 3, the graph for efficiency was much flatter than those for RMSEs and bias and demonstrated rapid change only in the extreme bins. These results suggested the better performance of item location estimates around 0, as assessed by RMSEs, was primarily due to bias and not efficiency.

One additional mixture condition. Based on the RMSE results from the extreme mixture condition, the local sample size hypothesis appeared incorrect. However, further analyses indicated these findings were not due to efficiency. To pursue these findings,

[3] In order to interpret results in a metric comparable to the previous conditions in which the person parameter distribution had a variance of one, the variance of the latent distribution in the mixture distributions was rescaled to one and the item parameter estimates were rescaled accordingly.

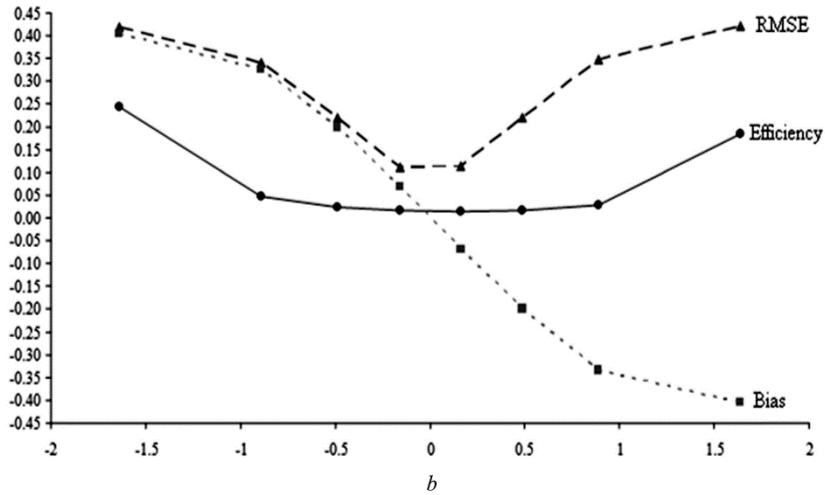

Figure 3:
RMSE, bias, and efficiency for a mixture distribution condition with sample size of 500, 8 items, and discrimination of 1.3. The underlying θ distribution was mixed normal such that 50% of the simulees were drawn from ~*N* (-2, .01) and 50% were drawn from ~*N* (2, .01). Markers are plotted at the mean of the true *b* parameters within each bin.

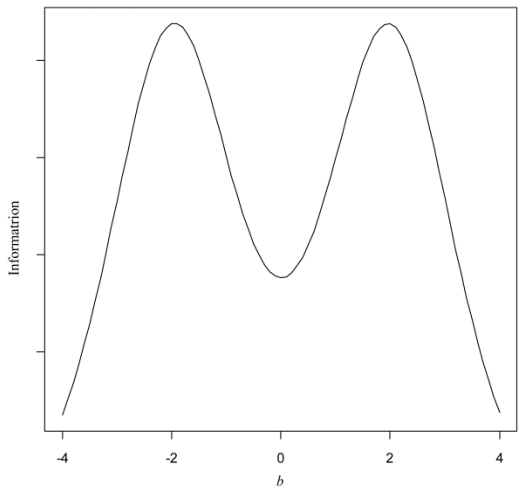

Figure 4:
Information function of mixed normal distribution with sample size of 500. The underlying θ distribution was mixed normal such that 50% of the simulees were drawn ~*N*(-2, .01) and 50% were drawn from ~*N*(2, .01).The items are assumed to have an $a = 1.3$. Information provided about the items is on the *y*-axis and the item location along the latent continuum is on the *x*-axis.

we investigated the information functions associated with the estimation of *b* parameters (e.g., Hambleton & Swaminathan, 1985) for various mixture distributions. The information function reflects the stability of estimated *b*s as a function of various θ values. Figure 4 provides an example of the information function for the extreme mixture condition reported in Figure 3. As shown in Figure 4, less information about items was provided by θ s of persons in the middle of the distribution in contrast to adjoining θ s. These results were similar to those for efficiency except that information increased to some extent at 0 rather than demonstrating a lack of change. This finding suggested that greater efficiency could be demonstrated in the middle of the distribution with a more sensitively designed condition.

Accordingly, a condition was created in which a ninth item was added with a *b* value of 0 to facilitate depiction of efficiency at this point of interest. Additionally, the item locations in this condition were fixed across replications at the locations corresponding to the means of the true *b* values for the 8 bins in the previous conditions: -1.64, -0.89, -0.49, -0.16, 0.16, 0.49, 0.89, and 1.64. As shown in Figure 5, efficiency was worse for the item located at 0 than for either of the immediately adjacent items as expected based on the person information functions.

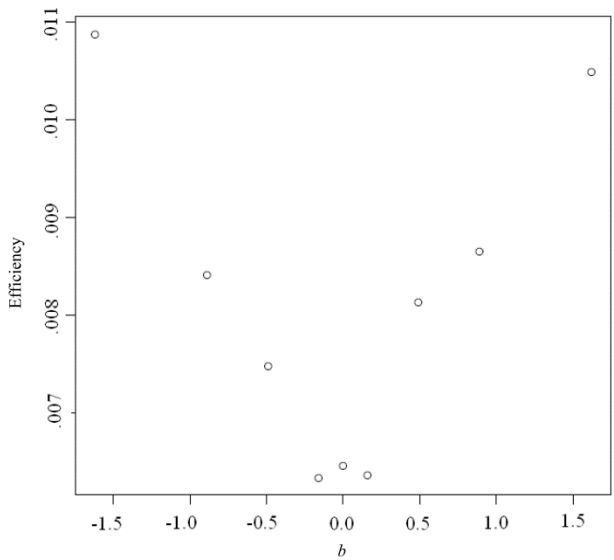

Figure 5:
Efficiency (*y*-axis) as a function of true item locations (*x*-axis) for mixed normal condition with sample size of 500, 9 items, and item discriminations of 1.3. The underlying θ distribution was mixed normal such that 50% of the simulees were drawn from a *N*(-2, .01) and 50% were drawn from a *N*(2, .01). Markers are plotted at the 9 item locations, which were held constant across replications of this condition.

Discussion

The current study examines the recovery of 1-PL person and item parameters in conditions that mimic small-scale testing applications. In general, our results are promising for practitioners wishing to use IRT in small-scale testing contexts where a mismatch between the items and persons distributions is anticipated.

Person parameter recovery

Estimates of θ were minimally affected by sample size and the relative alignment of person and item parameters. Stronger effects were observed for test length and item discrimination. Consistent with previous research simulating large-scale educational testing conditions, our study indicates that the precision (as measured by RMSE) of θ estimates is improved with longer tests. Even under non-optimal conditions (i.e., low discrimination, small sample size, and mismatched item-person distributions), the bias in θ estimates for simulees near the mean of the distribution is close to $\pm$ 0.10 logits. Although the situation worsens for θ estimates towards the ends of the distribution, the bias and precision of these estimates would be reasonable for small-scale testing situations.

In addition to test length, the accuracy of θ estimates was influenced by item discrimination; tests comprised of a set of more discriminating items yielded greater accuracy than those comprised of less discriminating items. Given the mathematics underlying the model and results from previous simulation studies, the inverse relationship between the discrimination parameters and RMSEs observed in our study was expected. Our simulations indicated this relationship was most pronounced at the extremes of the θ distribution, where estimates were most discrepant from the true θ values. In particular, RMSEs at the extremes were typically reduced by approximately 0.30 logits when the more discriminating set of items was used. On a more general note, item discrimination is often ignored when fitting a 1-PL model. However, for any two sets of items that separately conform to a 1-PL model, where one set contains more discriminating items than the other, the set with more discriminating items will yield more accurate θ estimates. This suggests that if the goal is to maximize the accuracy of θ estimates, item discrimination should not be ignored when fitting a 1-PL model; rather, the "most discriminating" and homogeneous set of items possible should be employed.

In general, small-scale test developers should be encouraged by our findings that tests as short as eight items and with samples as small as 100 participants generate reasonable estimates of person parameters, at least when operating under standards used in previous research (e.g., Hulin et al.,1982; Stone, 1992). We note, however, the acceptable level of precision with respect to θ estimates for any two researchers or for any two testing situations is likely to differ and depends on the context, the desired usage of the test, and the consequences of making decisions about people.

Item parameter recovery

Unlike the results for θ parameter recovery, the recovery of *b* parameters was strongly affected by sample size. Notably, however, item locations were estimated with greater precision than θ estimates; speaking generally of their average values across conditions, the lowest θ RMSEs approximated the highest *b* RMSEs. The deterioration of *b* parameter recovery follows a predictable pattern, depicted in Figure 6. As sample size increased, the recovery of *b* parameters improved. This reduction in error, however, was not linear with respect to the sample size. For example, adding 25 subjects to a sample size of 50 yielded larger decreases in RMSEs than adding 250 subjects to an already sizeable sample size of 750. In designing their studies, researchers can use our results for scales of 8 and 16 items (Figure 6) to approximate the cost/benefit of additional participants in terms of the diminishing returns in estimation accuracy.

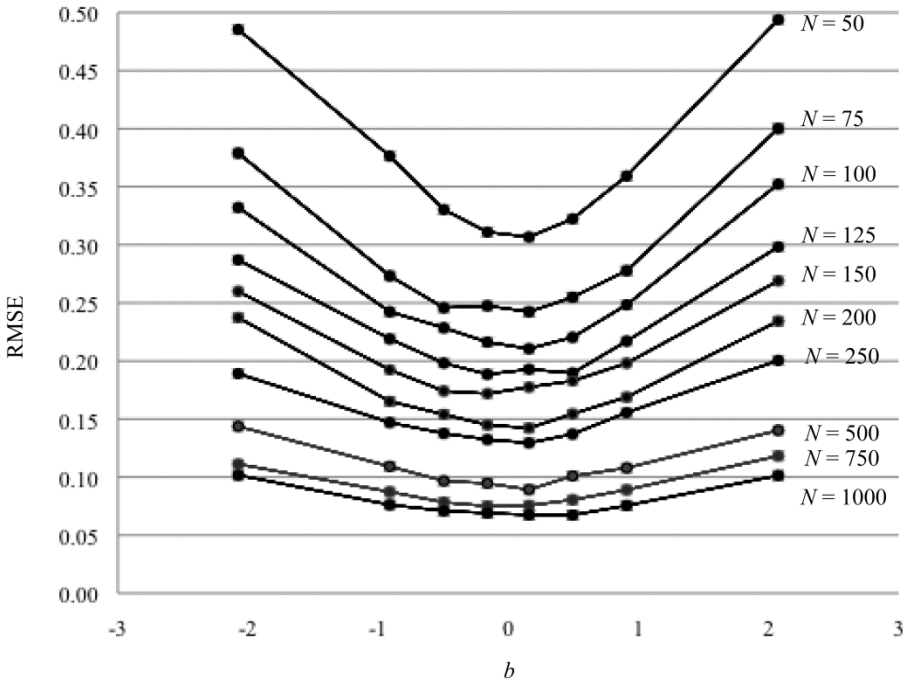

Figure 6:
Item location parameter RMSEs for various sample sizes for matched conditions with 8 items and discrimination of 1.3. Markers are plotted at bin midpoints. To approximate these RMSEs to their corresponding values when discrimination was .7, use $\hat{y} = 2.303x - .107$.

Item discrimination and test length also affected the recovery of *b* parameters. When items were generated with greater discrimination, the precision of *b* estimates was improved particularly for conditions with a sample size of 100. Though smaller than the other effects, test length improved *b* parameter recovery when sample size and discrimination were low. For small-scale testing situations, often the number of items is easier to increase than the number of participants; this may not be the case in large-scale situations in which item development is expensive. These effects are encouraging because item development provides an alternative or supplemental means of improving item parameter estimation other than increasing the number of participants.

The local sample size hypothesis. Based on the results of the original 16 conditions, we hypothesized that the precision of item location estimates depended in part on local sample size. Recall that for these conditions the alignment of item and person parameters was manipulated by shifting the item locations in the positive direction on the latent continuum. In doing so, the average distance between items and people was increased. Consistent with the local sample size hypothesis, the RMSEs of *b* estimates were lowest in the middle of the distribution in the match conditions. In addition, the estimation of *b* parameters at the low end of the scale improved while recovery of the *b* parameters at the high end deteriorated with mismatch conditions.

Additional conditions were constructed with mixture distributions of simulees to better understand the relationship between local sample size and the recovery of *b* parameters. In one of the mixture conditions, 20% of simulees were drawn from a $N(-.5, .0001)$ distribution and 80% were drawn from a $N(.5, .0001)$ distribution. Items located near the 80%-cluster of simulees were recovered as well as items closer to the center of the *b* distribution, but items located near the 20%-cluster of simulees were not, which constituted only mixed support for the local sample size hypothesis. The ambiguity of these findings motivated additional investigation.

The results of a more extreme condition with simulees highly localized at ±2 ran counter to the local sample size hypothesis in that items at the center of the item parameter distribution – where no simulees were located – were estimated best. Although these results were unexpected, the plot of information about items provided by θ parameters (Figure 4) followed an expected pattern; the most information was available where simulees were located, and a local minimum was reached at the center of the latent distribution. The information function in Figure 4 suggested that an item located exactly at zero on the latent continuum ought to be estimated with less precision than items located to either side of zero. A simulation confirmed this hypothesis (see Figure 5). Accordingly, the lower accuracy of estimation toward the extremes of the distribution was primarily due to bias (Figure 3). We speculate that nontrivial bias was observed in these conditions with mixed normal distributions of θ but not in the other conditions because a mixed normal distribution of θ departs from BILOG's default assumption of a standard normal distribution of θ. Future research on fitting the 1-PL assuming mixed normal distributions of θ would speak to this explanation.

With respect to the estimation of *b* parameters, our results indicated that it would take an extreme situation in practice to achieve sizable degradation of precision in the middle of

the latent distribution. Such an extreme situation, for example, would include $b \sim N(2, .01)$ and $\theta \sim N(-2, .01)$; estimation of item parameters would be severely compromised because there is an absence of information about items from θs. Put another way, if every item is so extreme that almost none of the people from a given population endorse the item, very little information is available from the item response data for estimating the model parameters.

Aside from situations such as this, our results should be encouraging for small-scale test developers; even for the small-scale conditions simulated in this study, item parameter estimates were reasonably accurate, with RMSEs ranging from .09 to .14 (for eight items, $N = 500$, and $a = 1.3$). Even for less ideal conditions ($N = 100$, eight items, $a = .7$, and mismatch of .5), bias was minimal (range from -.08 to .15) and RMSEs were considerably small, ranging from .53 to .74 at the extremes of the latent continuum. Estimation improved to the extent that sample size, item discrimination, and test length increased and mismatch decreased. As discussed above, in addition to the customary information functions provided for θ by items, we advocate the use of plots of item information provided by θ to better understand what persons bring to the estimation of item parameters before fitting the 1-PL, or any item response model, to observed data.

Widening our window

Our findings contribute to the literature in several ways. To our knowledge, no research has systematically investigated the relationship between alignment of θ and *b* distributions and the precision of item and person parameter estimation. Although our research indicates matching θ and *b* distributions is ideal, we also found a reasonable amount of mismatch has little effect on θ estimates and a fairly small effect on *b* estimates. That is, error in θ and *b* estimates may be sufficiently low despite some mismatch. We observed that when *b* parameters were located such that the θs fell on both sides of the *b*s, item recovery was better in the middle of the distribution, even though strong deviations from normality were present.

Through our supplemental conditions, we demonstrated additionally that the utility of item parameter information functions as a useful tool for researchers to indicate the anticipated utility of a set of items. In general, we believe these functions to be underutilized. Although such functions do not take into account the sampling error that will be present in applied studies, researchers can explore how different populations of subjects would potentially impact the information gained about a set of items. In this way, the effects of using various types of convenience samples may be evaluated relative to more comprehensive and expensive data-collection scenarios before any data are collected.

Although not included in the results section due to space considerations, the use of alternative prior distributions for the theta parameter was investigated for select conditions in order to make sure that our results were not idiosyncratic to our choice of the prior distribution. We ran a limited investigation to ensure that the results were reasonably generalizable to other priors. We limited the number of replications per condition to 10 in that

the analyses could be only partially automated; however, we believe that this number of replications was sufficient to assess generalization across alternative priors, particularly given that we evaluated these priors across a number of conditions. Specifically, the extreme conditions reported in Figures 2 and 3 were re-estimated using uniform, bimodal, and floating priors. Given that the default prior for the latent distribution in BILOG is standard normal, the generated data in these conditions did not match well with the prior (i.e., underlying theta distribution was bimodal). As such, the reported results for these conditions could potentially be improved upon by specifying a prior more consistent with the underlying data; in applied settings this would only be possible via domain expertise. For situations in practice where ability distributions are unknown, but suspected of departing dramatically from normality, researchers might wish to implement a uniform (ignorant) prior as part of a more conservative modeling approach.

We found that researchers using uniform priors could still expect to achieve parameter estimates of similar quality to those reported in the results section. In fact, the patterns of results using all three alternative prior specifications resulted in more accurate recovery of item location parameters marginally, with error reduction occurring most noticeably in the tails of the distributions. The item difficulty parameter RMSE plots were similar in shape to the reported results using default priors (see Figures 2 and 3), and were consistent with the item parameter information functions. However, for the conditions comprising the original design of this study (and for conditions most typical in practice), a normal prior remains an appropriate choice.

In addition, within the Rasch modeling literature, several alternatives have been proposed for dealing with small sample sizes. In particular, the characteristics of the Rasch family of models and the availability of sufficient statistics allow for CML as an alternative method for estimation. Further, a variety of proposed tests for data-model fit, such as those available in `eRm` package in R (Mair, Hatzinger, & Maier, 2012), are now available to researchers who work with the Rasch family of models (for further detail, see Hatzinger & Rusch, 2009; Koller & Hatzinger, 2012; Mair & Hatzinger, 2007a, 2007b; Ponocny, 2001). Further, non-parametric solutions to model-fit check have been proposed, and these solutions may be useful in conditions with small sample sizes. A series of estimation alternatives for Rasch models have been recently proposed (for a detailed discussion, see Ponocny, 2010). Also, with respect to the estimation procedures, it should be noted that CML is available only for estimation of Rasch models. This is similar to, although not identical to, our study of the 1-PL, given that we constrained discrimination parameters to be constant across items but not fixed to 1.00 (as is the case in the Rasch model). The results from the 1-PL studied here may be translated to a Rasch model (and vice versa) via a rescaling of the latent continuum.

Our purpose was to inform those wishing to use IRT methods in situations where the person and item location distributions do not match (e.g., potentially in small-scale testing environments), so there was an expectation of poor θ recovery relative to large-scale environments. We evidenced the level of error that researchers can expect when pushing the limits of sample size and test length requirements for IRT applications. Estimates that are relatively poor from a more traditional viewpoint may still be useful for broad classifications, such as for the first stage in multistage testing (Lord, 1980). From a substantive

standpoint, the level of error a researcher is willing to accept will directly impact the number of subjects and/or the number of items one should have for the analysis. The level of acceptable error is driven by the purpose of the research, such that for lower stakes testing or piloting, researchers might be more willing to recover estimates with more error in order to pursue IRT-based methods. We recommend using the results of information functions and simulations to make informed design choices within the context of a given research purpose.

Author note

This research was initiated when all the authors were members of the faculty-student research group in the Measurement, Statistics, and Methodological Studies at Arizona State University. Dubravka Svetina is now at Indiana University, Bloomington, IN, and Joanna S. Gorin is now at Educational Testing Services, Princeton, NJ.

References

Ankenmann, R. D., & Stone, C. A. (1992, April). *A Monte Carlo study of marginal maximum likelihood parameter estimates for the graded model.* Paper presented at the annual meeting of the National Council for Measurement in Education, San Francisco, CA.

Baldwin, S., Baldwin, P., & Nering, M. (2007, April). *A Comparison of IRT equating methods on recovering item parameters and growth in mixed-format tests.* Paper presented at the annual meeting of American Educational Research Association, Chicago, IL.

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431-444. doi: 10.1177/014662168200600405

Chen, S., Hou, L., Fitzpatrick, S. J., & Dodd, B. G. (1997). The effect of population distribution and methods of theta estimation on CAT using the rating scale model. *Educational and Psychological Measurement, 57,* 422-439. doi: 10.1177/0013164497057003004

Childs, R. A., & Chen, W.-H. (1999). Software Note: Obtaining comparable item parameters estimates in MULTILOG and PARSCALE for two polytomous IRT models. *Applied Psychological Measurement, 23,* 371-379. doi: 10.1177/01466219922031482

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement, 53,* 61-77. doi: 10.1177/0013164493053001005

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13,* 77-90. doi: 10.1177/014662168901300108

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests [Introduction to Mental Test Theory]*. Huber, Bern. In German.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S., & Sheih, Y. Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions, and item pool characteristics. *Applied Psychological Measurement, 29*, 433-456. doi: 10.1177/0146621605280072

Hambleton, R. K. (1989). Principles and selected applications of items response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). New York: Macmillian.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff Publishing.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*, 279-291. doi: 10.1177/014662169101500308

Hatzinger, R., & Rusch, T. (2009) IRT models with relaxed assumptions in eRm: A manual-like instruction. *Psychology Science Quarterly, 51*, 87–120.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260. doi: 10.1177/014662168200600301

Jurich, D., & Goodman, J. (2009, October). *A Comparison of IRT parameter recovery in mixed format examinations using PARSCALE and ICL*. Paper presented at the Annual meeting of Northeastern Educational Research Association, Rocky Hill, CT.

Kim, S. H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26*, 25-41. doi: 10.1177/0146621602026001002

Koller, I. (2010). Item response models in practice: Testing the Rasch model in small samples and comparing different models for measuring change. Unpublished dissertation. Alpen-Adria-Universität Klagenfurt.

Koller, I. & Hatzinger, R. (2012, April). Exakte Tests für das Rasch Modell unter besonderer Berücksichtigung von lokaler stochastischer Unabhängigkeit [Exact tests for the Rasch model with special consideration of local stochastic independence]. In I. Koller, M. J. Maier, K. Gruber, R. W. Alexandrowicz, & I. W. Nader. *Item Response Modelle: Weiterentwicklung, Überprüfung und Anwendung* [*Item Response Models: Development, Testing, and Application*]. Symposium gehalten bei der 10. Tagung der Österreichischen Gesellschaft für Psychologie, Graz. [Symposium held at the 10th Meeting of the Austrian Society for Psychology, Graz.] In German.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*, 328.

Linacre, J. M. (2000). WINSTEPS: Rasch analysis for all two-facet models, version 3.04. Chicago: MESA Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1983). Small N justifies Rasch model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 51-61). New York: Academic Press.

Mair, P., & Hatzinger, R. (2007a). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science, 49*, 26-43

Mair, P., & Hatzinger, R. (2007b). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*, 1-20.

Mair, P., & Hatzinger, R., Maier, M. J. (2012). eRm: Extended Rasch modeling. R package version 0.15-1. URL: http://cran.r-project.org/web/packages/eRm/index.html

Mislevy, R. J., & Bock, R. D. (1982). Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program. In: Item Response Theory and Computerized Adaptive Testing Conference Proceedings. Wayzata, MN.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57-75. doi: 10.1177/014662168901300106

Partchev, I. (2010). `irtoys`: Simple interface to the estimation and plotting of IRT models. R package version 0.1.3. URL: http://cran.r-project.org/web/packages/irtoys/index.html

Ponocny, I. (2001). Nonparametric goodness-of-fit tests for the Rasch model. *Psychometrika, 66*, 437-460.

R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.

Seong, T. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*, 299-311. doi: 10.1177/014662169001400307

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*, 1-16. doi: 10.1177/014662169201600101

Wang, W.-C., & Chen, C.-T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program for the family of Rasch models. *Educational and Psychological Measurement, 65*, 376-404. DOI: 10.1177/0013164404268673

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. -S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 26*, 339-352. doi: 10.1177/0146621602026003007

Wright, B. D., & Douglas, G. A. (1975). Best test design and self-tailored testing. *Research Memorandum No. 19*, Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D., & Douglas, G. A. (1976). Rasch item analysis by hand. *Research Memorandum No. 21*, Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design: Rasch measurement.* Chicago: MESA Press.

Xiong, X., Lewis, C., & Mingmei, W. (2009, July). *Accuracy of estimating item parameters for the Rasch model based on small sample size*. Paper presented at the annual meeting of the Psychometric Society, Cambridge, England.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*, 275-291.