During the submission process, you will be providing information on the following specific items.

1. Title of the paper or symposium.

2. Desired session format (paper, symposium, roundtable, graduate student poster, teacher inquiry).

3. First author's name and institutional affiliation.

4. First author's e-mail address, USPS address, and telephone number.

5. Names of co-authors or co-presenters (if co-presenters will address different topics, such as in a

symposium, list the topic for each co-presenter) and their institutional affiliations.

6. If proposal is for a symposium, list the names and institutional affiliations of the organizer(s) and

facilitator of the symposium.

7. Abstract of the proposal. It should not exceed 100 words.

8. Proposal Summary submitted as a file upload. The main body of the proposal text should not

exceed five pages, double spaced. DO NOT IDENTIFY THE AUTHOR(S) OR THEIR

INSTITUTIONAL AFFILIATIONS ANYWHERE IN THE SUMMARY. This will provide for a

blind review of the proposal. This file should include:

a. Title of the paper or symposium

b. Purpose

c. Rationale and background

d. Data source (population, sample, setting)

e. Methods

f. Results and conclusions\*

g. Limitations

h. Implications

**TITLE (original):** Examining the Robustness and Comparability of PROX estimation with known item parameters.

Title (alternate): Comparing the Robustness of PROX Estimation to Maximum Likelihood Estimators with Fixed Item Parameters

**ABSTRACT**

OPITION 1: This study examines the robustness and comparability of a non-iterative estimation method (PROX) with more common estimation methods when item parameters are known. A simulation was conducted to manipulate the person ability distribution, item difficulty distribution, and sample size. Comparisons between PROX, JMLE, CMLE, and EAP recovery of person parameters were made in terms of bias, MAD, and RMSE. Results indicate that in typical conditions PROX is at least as good as other methods, and in more extreme conditions performs worse than only JMLE for person parameter recovery. Future avenues of investigation are also discussed.

OPTION 2: This study investigates the effects of person ability distribution, item difficulty distribution, and sample size on the robustness and comparability of a non-iterative estimation method (PROX) with more common estimation methods when item parameters are known.

**PURPOSE**

Iterative estimation procedures have been universally adopted for Item Response Theory and Rasch Measurement Theory software packages. These iterative methods are used to produce estimates of person ability even in situations where item difficulty is known, such as exams delivered via calibrated item banks. Under a Rasch Measurement paradigm, it may be argued that in situations where all item parameters are fixed, determining person ability is more calculation than estimation.

Cohen (1979) introduced PROX, a non-iterative method for estimating Rasch measures when data are complete and both items and persons are approximately normally distributed. A method for hand calculating PROX is shown in Wright and Stone (1979, CH 2). Although not widely applied as a stand-alone estimation method, non-iterative PROX does provide the initial estimates for Joint Maximum Likelihood Estimation in Winsteps. Linacre (1994) describes iterative PROX estimation equations for missing data. When item parameters are fixed, this iterative version of PROX becomes non-iterative.

This study examines three primary research questions:

1. How robust is non-iterative PROX to violations of the distributional assumptions for items and persons?
2. How robust is non-iterative PROX to sample size fluctuations?
3. How do estimates produced by non-iterative PROX compare to other common estimation methods under the conditions set in the first two research questions?

**BACKGROUND**

In the certification and licensure industries, there are multiple considerations to be made at each step of creating, updating, or even maintaining an examination program. There is much written about these steps (see the Handbook of Test Development for an overview), but for the estimation of person ability parameters there is very little investigation into how the PROX estimation procedure compares to more computationally intensive estimation procedures.

When PROX was developed, Wright & Douglas (1975) noted that the results from the method are “equivalent to estimates obtained from UCON [an iterative maximum-likelihood solution] for all practical purposes,” and Cohen’s (1979) own person ability estimates showed essentially equivalent estimates. However, estimation techniques have drastically improved since then, as has computational power. A review of the literature did not turn up any study comparing the non-iterative PROX to other iterative MLE methods.

**METHODS**

Using R, simulated Rasch items and person responses were created that adhere to the following conditions:

1. Normally distributed item difficulties and person abilities with μ = 0 and σ = 1 (“Standard Normal Parameters”)
2. Normally distributed item difficulties and person abilities with μ = 0 and σ = 2 (“Wide Normal Parameters”)
3. Normally distributed item difficulties (μ = -1, σ = 1) and person abilities (μ = +1 and σ = 1) (“Small Parameter Mismatch”; 2 logit difference)
4. Normally distributed item difficulties (μ = -2, σ = 1) and person abilities (μ = +2 and σ = 1) (“Large Parameter Mismatch”; 4 logit difference)
5. Normally distributed item difficulties (μ = -3, σ = 1) and person abilities (μ = +3 and σ = 1) (“Extreme Parameter Mismatch”; 6 logit difference)
6. Normally distributed item difficulties (μ = 0, σ = 1) and bimodally distributed person abilities (μ1 = -1.5, μ2 = +1.5, and σ1 = σ2 = 1) (“Bimodal Person Parameters”).

These different distribution combinations allow us to investigate the effects of distribution changes between item difficulties and person abilities among the different estimation methods.

The number of simulated persons will be one of n = 25, 50, 100, 250, 500, or 1000, and the number of items is fixed at 200. Given that we are simulating using the Rasch model, these numbers of persons reflect both a smallest viable sample (25) and a more than adequate sample (1000) to be confident in the calculation of person ability estimates in typical use cases. The choice of using 200 items for our simulation allows us to mimic the typical test lengths seen in certification and licensure examinations. Fully crossing the sample size and the distribution conditions leads to 36 simulated conditions, each of which were repeated 100 times for a total of 3600 conditions for analysis.

In addition to the PROX method, we compare three additional, common IRT estimators as implemented in specific R packages (shown in parentheses). These are as follows:

1. Joint Maximum Likelihood Estimation (JMLE; TAM package)
2. Conditional Maximum Likelihood Estimation (CMLE; eRm package)
3. Expected A Posteriori Estimation (EAP; ltm via the irtoys package)

The implementations of JMLE and EAP allow for fixed item parameters, so in these cases we provide the simulated item parameters to the estimation functions. However, for CMLE, we are unable to do this so item and person parameters are simultaneously estimated in this case.

For each estimation procedure, person ability estimates were forced to be between -4 and 4. Estimates that fall outside of this range (e.g., all 0 response patterns and all 1 response patterns) were coerced to the appropriate extreme per standard practice.

Person abilities estimated from each of the four procedures are compared to the true simulated ability estimates, creating estimates for mean bias, mean absolute error (MAD), and root mean square error (RMSE) for each iteration within a condition. These mean estimates within an iteration are then averaged across all iterations such that each condition has an averaged mean estimate of these measures.

Mean bias is a measure of the systematic error introduced by the estimation process and was calculated as:

MAD is a measure of all error in the estimation process, systematic and random, and was calculated as:

RMSE is another measure of all error in the estimation process but is equal to the standard deviation of the parameter when the process is unbiased. It was calculated as:

The formulas and definitions for these measures were taken from Feinberg & Rubright (2016).

Additionally, the correlation between the procedures averaged across the iterations are investigated to ensure consistency of estimation and check for any noticeable errors in the estimation procedures. This was calculated by creating a correlation matrix from all ability estimates in each of the replications, then averaging the resulting correlation matrix across the 100 replications to produce a mean correlation matrix for each of the 36 conditions.

**RESULTS**

**Mean Correlations**

The mean correlation between estimation procedures themselves and between the procedures and the true theta remained high (.96 or greater) for all conditions except the “Large” and “Extreme Parameter Mismatch” conditions. In these conditions, the estimation procedures maintained a high mean correlation between each other (.97 or greater) but had a significantly lower mean correlation to the true theta (approximately .88 in the “Large Parameter Mismatch” conditions and .68 in the “Extreme Parameter Mismatch” conditions). There was little variability in any of the mean correlations due to sample sizes.

**Average Mean Bias**

Averaged across all conditions, the JMLE estimator had the lowest average mean bias (-0.08), followed by the PROX estimator (0.14). The EAP estimator (-0.34) and the CMLE estimator (0.74) had much higher average mean bias. Table 1 shows the mean bias for each estimator across all conditions.

The estimators as a whole produce unbiased estimates when the difference between the item and person parameter means is negligible. Starting with the Small Parameter Mismatch (two logit difference in the means of the parameters), the CMLE starts producing extremely biased results, with an average mean bias of 1.02 across all sample sizes. EAP starts producing very slightly biased estimates of person parameters at this point as well.

Both CMLE and EAP produce even more biased estimates within the Larger Parameter Mismatch conditions (average mean bias: 1.82 and -0.41, respectively). The CMLE estimates become more biased as sample size increases, while EAP estimates become less biased as sample size increases. PROX estimates are slightly biased here as well (average mean bias: 0.23) but are unaffected by sample size. JMLE estimates are unbiased on average but are affected by sample size: slightly negatively biased with small sample sizes and slightly positively biased with large sample sizes.

**Average MAD and RMSE**

On average, PROX and JMLE had the lowest mean absolute difference across all conditions (0.31 and 0.33, respectively). EAP had a somewhat higher average MAD (0.47) while CMLE had a still-higher average MAD (0.67). Table 2 shows the average MAD values for each estimator across all conditions.

PROX and JMLE also had very similar average MAD values in most conditions. The noticeable exceptions are in the Wide Normal condition, where PROX had the highest average MAD values, and in the Extreme Parameter Mismatch condition, where PROX had the lowest average MAD values.

The MAD values for PROX were unaffected by sample size. The MAD values for JMLE and EAP were affected by sample size in the Large and Extreme Parameter Differences conditions, where larger sample sizes reduced the value of MAD. CMLE was also affected by sample size in these conditions, but its MAD values increased with sample size instead.

RMSE values are shown in Table 3 and follow the same trends for each estimator as MAD.

**CONCLUSIONS**

Different estimation methods have their own unique set of pros and cons, which analysts must consider depending on the testing situation.

Although non-iterative PROX has been shown to work well when the distributional requirements are met, the full impact of violations of these requirements is unknown. Results from this study show that in cases with even small mismatches (about 2 logits) between the average difficulty of the known Rasch items and the average ability of the examinees PROX continues to work essentially as well as JMLE in recovering person ability scores. Both estimators were essentially unbiased in these conditions, though PROX did have a larger variance in the estimates as compared to JMLE.

With greater parameter mismatch, both estimators performed worse but PROX was affected to a larger degree. From a practical perspective, this may not be a large concern as it would be very apparent from typical test development processes that the mismatch exists and would be addressed prior to delivering the exam in a high-stakes situation. If it is unavoidable, however, it is clearly better to use JMLE over PROX to recover person parameters when a mismatch between known items and unknown persons is greater than 2 logits.

The results also indicated that CMLE and EAP are about equivalent to PROX and JMLE when there is no more than a small mismatch between known item and person abilities, but these second set of estimators quickly introduce more error than the first set and should probably be avoided in those situations.

**LIMITATIONS AND IMPLICATIONS**

There are no R packages that estimate CMLE with fixed item parameters, so additional care must be taken with the interpretating the comparison of those estimates with other methods. There is also the possibility of different implementations of the estimating procedures in different programs and this study did not test an exhaustive number of procedures. Replicating these results with different programs and more procedures may strengthen the validity of these findings.

The decision to fix the number of items at 200 was made to ensure that the scope of the project was reasonable. Future research should consider the impact of test length on person estimates and how test length may interact with sample size as there was not a significant impact of sample size found in this study.

This study also utilized known item parameters to mimic a mature testing program. While this implementation of PROX requires known item parameters, future research may want to investigate how well PROX can recover person parameters if a separate process is used to estimate item parameters beforehand.

Finally, there were decisions to “correct” person parameter estimates for extreme values. While not likely to have a major effect on the results in most conditions, the Extreme Mismatch Condition was most likely to have been affected by this decision. A further investigation into the data from this condition is needed before stronger conclusions about the interaction of extreme item-person mismatch and ability estimation procedure.