Online Appendix B

In the main paper, we demonstrated that traditional item and model fit assessment does not necessarily detect 3PL item parameters that produce biased trait estimates. To further explore this surprising result, we considered whether 3PL trait estimation bias might be remedied by item elimination based on the $S-X^2$ item fit statistic.

We first investigated properties of the 3PL items calibrated with $N = 20,000$ that consistently did not fit across the 100 replications (i.e., across 100 sets of estimated item parameters). Noting that each data set was generated according to the same set of 100 4PL items, we noticed that certain items consistently misfit according to the $S-X^2$ statistic. Specifically, 17 items fit on at most 22% of replications, and the remaining 83 items fit on at least 70% of replications.[1] To assess whether $S-X^2$-based item elimination would also eliminate the bias in estimated trait scores, we re-estimated EAP2 latent trait scores for the 3PL by computing trait estimates from only those items that fit according to the $S-X^2$ statistic and $\alpha = .05$. We also estimated 3PL item parameters at sample sizes $N = 1,000$ and $N = 5,000$ and calculated EAP2 trait estimates both for the full set of 100 items and after eliminating misfitting items according to the $S-X^2$ statistic. First, we noticed striking differences in model and item fit across calibration sample sizes. According to the M_2 model fit index with $\alpha = .05$, 98% of 3PL models at both the $N = 1,000$ and $N = 5,000$ sample sizes fit; this is in contrast to the 81% of 3PL models that fit at the $N = 20,000$ sample size. According to the $S-X^2$ item elimination rule, between 87 and 98 (average 93) items were retained at sample size $N = 1,000$, between 77 and 92 (average 86) items were retained at sample size $N = 5,000$ and between 71 and 83 (average 76) items were retained at sample size $N = 20,000$. The resulting biases of latent trait estimates for the full item banks and for the misfit-eliminated item banks are displayed in Figure B1. This figure shows that at $N = 20,000$, item elimination based on $S-X^2$ largely eliminates the bias induced by using a misspecified IRT model. However, at the two smaller sizes, the trait estimation bias for the reduced item bank is nearly the same as for the full item bank. These results suggest that the $S-X^2$ statistic is not powerful enough to detect this type of functional form misfit in samples as large as $N = 5,000$. It may be that item elimination based on a different item fit measure or a stricter $S-X^2$ α criterion might result in less biased trait estimates for misspecified models and moderate to small calibration samples. Such explorations are beyond the scope of this paper, but may prove useful in reducing the effects of model misspecification in real data.

[1] We also noticed that these 17 items tended to have low data-generating b parameters and high data-generating a parameters, but not necessarily low data-generating d parameters (as one might expect). Thus, it is not necessarily items with d paramemters far below 1 that tend to not fit the 3PL, it is easy items that are highly discriminating. This may be the case because items that are easy and highly discriminating reach their upper asymptote at lower θ values than do other combinations of 4PL item parameters.

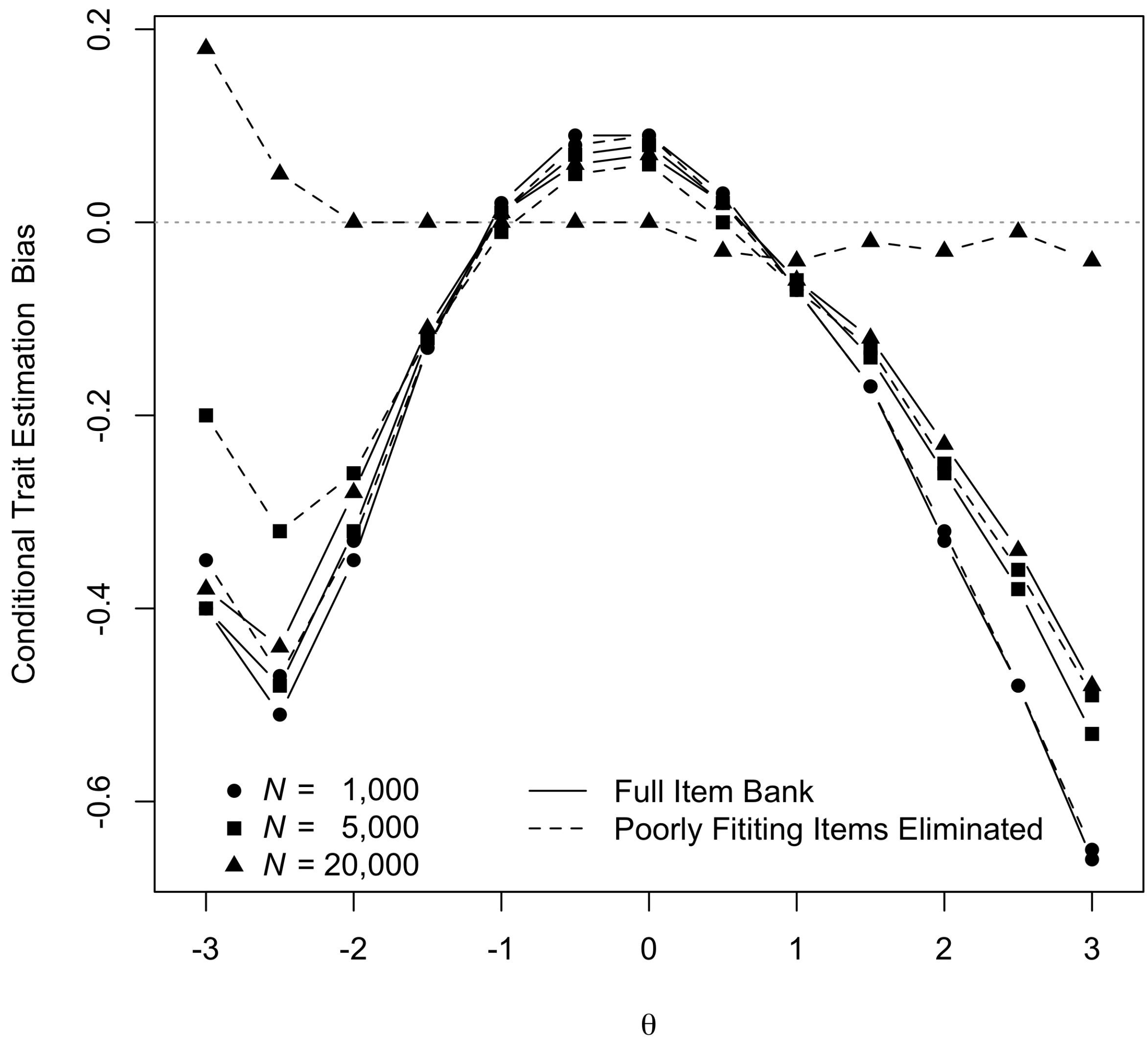

Figure B1. Bias of 3PL EAP2 trait estimates before and after item elimination and at three item calibration sample sizes ($N = 1,000$, $N = 5,000$, and $N = 20,000$). In this figure, all data were generated under the 4PL, item parameters were estimated with the 3PL, and trait estimates were computed from 4PL-generated item responses and 3PL item parameter estimates. In the "full item bank" condition, trait estimates were computed from the full set of 100 items. In the "poorly fitting items eliminated" condition, items that did not fit the model according to the $S-X^2$ item fit statistic and $\alpha = .05$ were not used to estimate trait scores.