Item-Weighted Expected a Posteriori Method for Improved Latent Trait Estimation in Item Response Theory

Udi Alter & R. Philip Chalmers

York University – Université York
Quantitative Methods, Dept. of Psychology · York University

This study proposes a novel method for estimating respondents' abilities using item response theory (IRT) models. The proposed technique extends the expected a posteriori (EAP) estimation method (Bock & Aitkin, 1981) by incorporating a standardized weight function based on either user-defined values or item-fit statistics. The standardized weight values range from 0 to 1, where responses from items with lower weight values contribute less to the ability estimates. We used a Monte Carlo simulation to evaluate the new item-weighted expected a posteriori (IWEAP) approach and compare it to the common ability estimation technique.

Item-Weighted Expected a Posterior (IWEAP)

IWEAP likelihood function for a dichotomous item

$$P(\mathbf{y}|\theta,\mathbf{w}) = \prod_{j=1}^{J} [P_j(y=1|X_q)^{y_j} \cdot (1 - P_j(y=1|X_q))^{1-y_j}]^{w_j}$$

$\mathbf{w} = [w_1, w_2, \ldots, w_j]$, where $\forall w \geq 0$ and $\forall w \leq 1$

Where the w_j terms are the user-define input weights for item j:

- $w_j = 1$ leads to a standard, full weight for the item
- $w_j < 1$ indicates less weight
- $w_j = 0$ the item is omitted entirely from the likelihood

Objective

Evaluate how well the new IWEAP approach estimates person ability scores and compare it to the technique across different sample sizes, test lengths, and proportions of item-misfit

Method

$S\text{-}X^2$ (Orlando & Thissen, 2000) item-fit statistic was used as weight

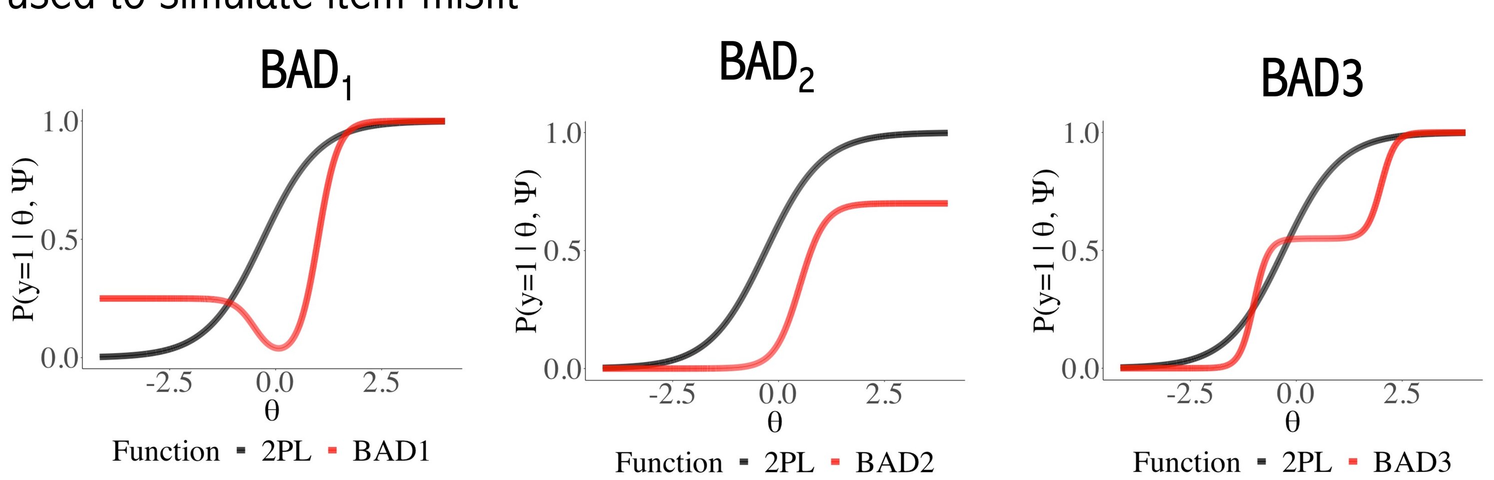
$$\text{IWEAP: } w_j = 1 \Big/ \sqrt{\frac{(s-X^2)_j}{DF_j}}$$

Monte Carlo Simulation

5000 simulations were conducted per condition. For each condition, a 2PL IRT model with dichotomous responses

- Sample size: N = 250, 500, 1000
- Test length: 10, 20, 40
- Proportion of bad-fitting items: 0%, 10%, 20%

Three types of atypical IRT models (red) from Orlando & Thissen (2003) were used to simulate item-misfit



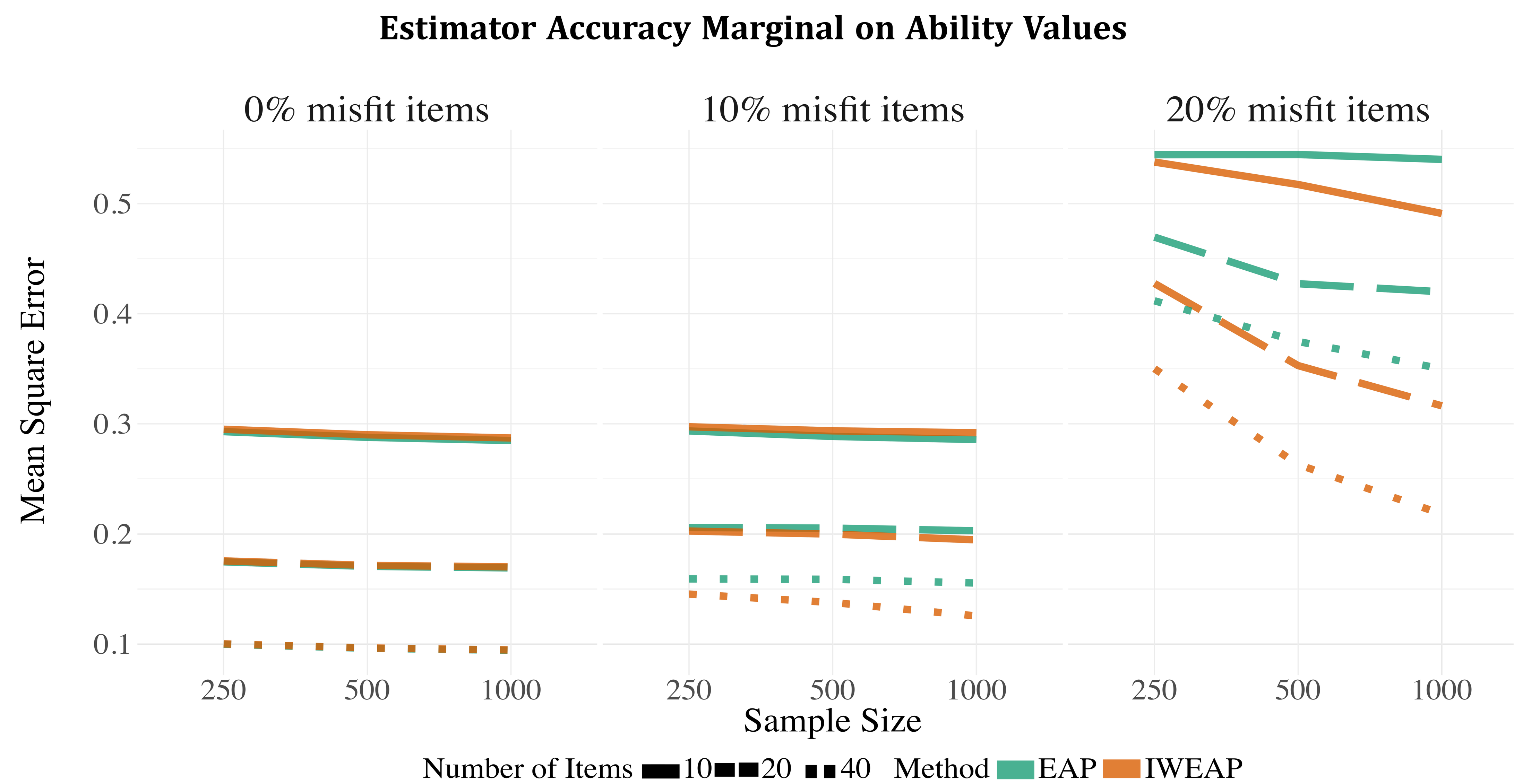


Figure 1. Simulation results comparing expected a posteriori (EAP; green) with the proposed item-weighted expected a posteriori (IWEAP; in orange). Results are faceted by the percent of misfitted items in the 2PL IRT model with dichotomous responses. A solid line indicates the model includes 10 items, a long-dashed line 20 items, and a dashed line 40 items.

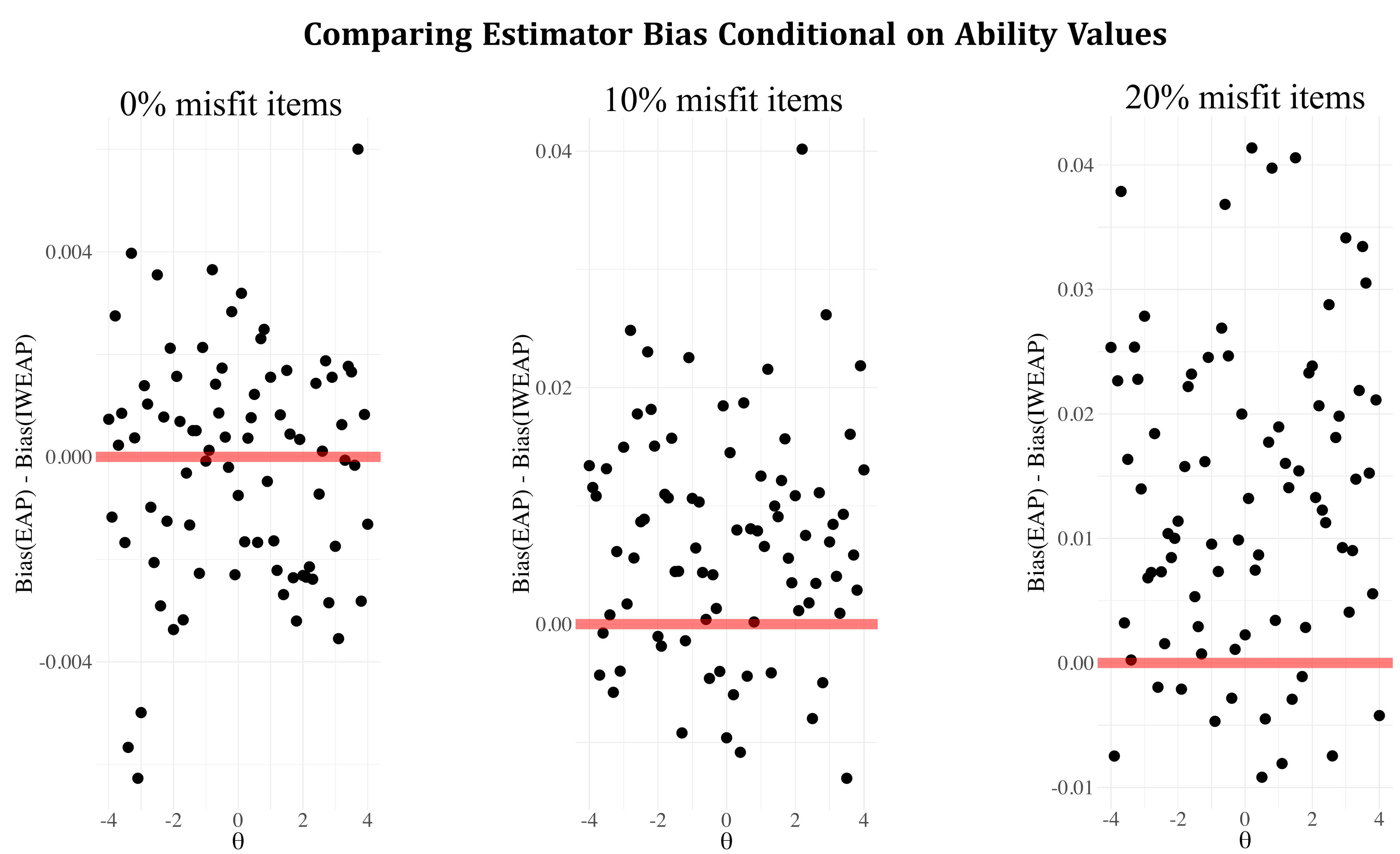


Figure 2. Simulation results comparing the bias of expected a posteriori (EAP) and item-weighted expected a posteriori (IWEAP) across the ability range faceted by the level of misfit. In these plots, N=1000, test length = 40, and the type of misfit is BAD1 (for the 10% and 20% misfit conditions). Observations above the red line (i.e., 0) indicate that the IWEAP bias is smaller than that of EAP.

Conclusions

No Misfit
As expected, EAP and IWEAP performed virtually the same regardless of sample size and test length in the no misfit conditions.

Little Misfit (10% of items)
The difference between the estimators was negligible across all sample sizes. That difference increases with longer tests where IWEAP yielded slightly less biased estimates than EAP.

Greater Misfit (20% of items)
Ability estimates from EAP were the most biased, especially with short test lengths and smaller sample sizes. Conversely, IWEAP performed considerably better when there are misfitting items present.

Ability estimates from IWEAP were consistently more accurate than those from EAP across all test lengths and sample sizes.

Discussion and Future Directions

IWEAP demonstrated improved ability estimation over and above EAP when item-misfit was present with substantially better accuracy among longer tests, larger sample sizes, and greater misfit.

IWEAP estimation should be tested using IRT models with polytomous response options and various item-fit statistics (e.g., Stone's χ^{2*}, 2000)

Scan for code and data