Estimating Treatment Effects with the Explanatory Item Response Model

Journal of
Research on
Educational Effectiveness

Joshua Gilbert

How Much do Scoring Methods Matter for Causal Inference?

The manner in which student outcomes are measured, scored, and analyzed often receives too little attention in randomized experiments. In this study, we aimed to explore the consequences of different scoring approaches for causal inference on test score data. We compared the performance of four methods, Classical Test Theory (CTT) sum scores, CTT mean scores, item response theory (IRT) scores, and the Explanatory Item Response Model (EIRM). In contrast to the CTT- and IRT-based approaches that score the test and estimate treatment effects in two separate steps, the EIRM is a latent variable model that allows for simultaneous estimation of student ability and the treatment effect. The EIRM has a long history in psychometric research, but applications to empirical causal inference settings are rare. Our results show that which model performs best depends on the context.

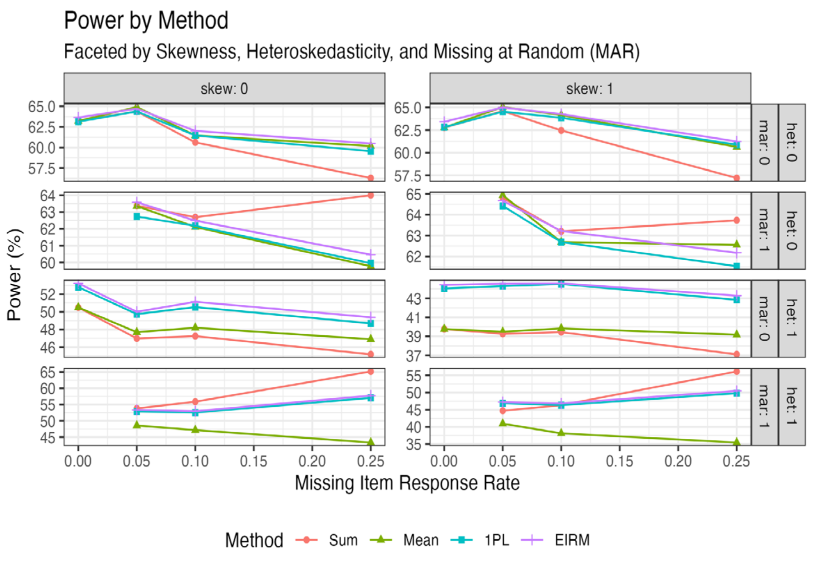

Chart notes: Statistical power (y-axis) by missing item response rate (x-axis) and estimation method (color and shape) shows that the relative performance of each approach depends on the context. The EIRM and IRT-based scores are more robust to missing data and provide the most benefits to power when the latent trait is heteroskedastic. Legend: skew = latent trait is skewed, het = latent trait is heteroskedastic, mar = item responses are missing at random, sum = sum score, mean = mean score, 1PL = IRT theta score, EIRM = explanatory item response model.

Comparative Model Performance

To determine the conditions under which each scoring model was most effective, we conducted a Monte Carlo simulation study that examined the performance of CTT-based sum scores, IRT-based scores, and the EIRM across a range of conditions, including the rate of missing item response data, missingness mechanism, heteroskedasticity, and skewness. We found that bias and false positive rates were similar across all conditions, though IRT-based scores and the EIRM provided superior calibration of standard errors under model misspecification. In terms of statistical power, the EIRM and IRT-based scores were more robust to missing item response data than other methods when parametric assumptions are met and provided a moderate benefit to statistical power under heteroskedasticity, but their performance was mixed under other conditions. We concluded by applying the various scoring methods to empirical data from a reading comprehension assessment and found that the EIRM provided a moderately more powerful and precise estimate of treatment impact.

What Scoring Method Should You Use? It Depends

In summary, our results indicate that there is no single model that excels across all metrics in all circumstances. Instead, performance relies on the tenability of parametric assumptions. While the advantages of the EIRM appear to be modest, it is nonetheless a potentially useful tool for the applied researcher examining causal effects on test score data, particularly in the presence of missing data.

Full Article Citation:
Gilbert (2024). Estimating Treatment Effects with the Explanatory Item Response Model. *Journal of Research on Educational Effectiveness*, Volume(Issue), pages, DOI: https://doi.org/10.1080/19345747.2023.2287601.