EdWorkingPaper No. 22-677

Estimating Treatment Effects with the Explanatory Item Response Model

Joshua B. Gilbert
Harvard University

This simulation study examines the characteristics of the Explanatory Item Response Model (EIRM) when estimating treatment effects when compared to classical test theory (CTT) sum and mean scores and item response theory (IRT)-based theta scores. Results show that the EIRM and IRT theta scores provide generally equivalent bias and false positive rates compared to CTT scores and superior calibration of standard errors under model misspecification. Analysis of the statistical power of each method reveals that the EIRM and IRT theta scores provide a marginal benefit to power and are more robust to missing data than other methods when parametric assumptions are met and provide a substantial benefit to power under heteroskedasticity, but their performance is mixed under other conditions. The methods are illustrated with an empirical data application examining the causal effect of an elementary school literacy intervention on reading comprehension test scores and demonstrates that the EIRM provides a more precise estimate of the average treatment effect than the CTT or IRT theta score approaches. Tradeoffs of model selection and interpretation are discussed.

VERSION: November 2022

Suggested citation: Gilbert, Joshua B.. (2022). Estimating Treatment Effects with the Explanatory Item Response Model. (EdWorkingPaper: 22-677). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/snvz-ew19

Estimating Treatment Effects with the Explanatory Item Response Model

Joshua B. Gilbert

Harvard University Graduate School of Education

joshua_gilbert@g.harvard.edu

Abstract

This simulation study examines the characteristics of the Explanatory Item Response Model (EIRM) when estimating treatment effects when compared to classical test theory (CTT) sum and mean scores and item response theory (IRT)-based theta scores. Results show that the EIRM and IRT theta scores provide generally equivalent bias and false positive rates compared to CTT scores and superior calibration of standard errors under model misspecification. Analysis of the statistical power of each method reveals that the EIRM and IRT theta scores provide a marginal benefit to power and are more robust to missing data than other methods when parametric assumptions are met and provide a substantial benefit to power under heteroskedasticity, but their performance is mixed under other conditions. The methods are illustrated with an empirical data application examining the causal effect of an elementary school literacy intervention on reading comprehension test scores and demonstrates that the EIRM provides a more precise estimate of the average treatment effect than the CTT or IRT theta score approaches. Tradeoffs of model selection and interpretation are discussed.

Keywords: Explanatory Item Response Model, causal inference, statistical power, simulation, educational measurement

Estimating Treatment Effects with the Explanatory Item Response Model

When estimating treatment effects in education research, test scores are usually analyzed as a single value, calculated from the responses to individual items. Typically, a total correct sum score, mean score, or an Item Response Theory (IRT)-based theta score is used. As proposed by Wilson and De Boeck (2004), the Explanatory Item Response Model (EIRM), which models the item responses directly rather than as single summary value, provides an alternative approach to modeling assessment data. The EIRM has theoretical appeal because it capitalizes on all available item response information without the need to reduce the multiple item responses to a single score for subsequent analysis, which can result in biased parameter estimates due to measurement error (Ye, 2016, p. 43-44). Similarly, Briggs (2008) argued that a key benefit of the EIRM is the ability to combine psychometric measurement and explanatory regression analysis into a single model, in contrast to the more traditional two-step process, in which the measurement and explanatory models are independent of one another. Past studies have employed the EIRM when the research question is concerned with relationships between fixed person and item characteristics and item response patterns (see for example Hartig et al. 2011; Briggs 2008; Randall, Cheong, & Engelhard 2011), or the theoretical benefits of modeling item responses directly (Zwinderman, 1991; Christensen, 2006). However, in the causal inference context, relatively few studies have employed the EIRM when estimating the treatment effects of educational interventions (see for example Kim, et al., 2022 for a cluster-randomized trial; Rabbitt, 2018 in an instrumental variables context; Stevenson, et al., 2013 for a pre-post study), and we are aware of no studies to date that have empirically examined the potential benefits or tradeoffs of employing the EIRM in causal inference contexts.

The purpose of this study is twofold. First, to determine to what extent the theoretical advantages of the EIRM obtain under various simulated conditions of sample size, number of items, treatment effect size, the rate of missing item response data, and to what extent differences in model performance depend on model misspecification. Second, to determine to what extent the results of the simulation are consistent with the application of the EIRM to the estimation of treatment effects in an example of empirical educational assessment data. The effects of missing item response data are emphasized because missing item responses are treated in contrasting ways across the different models and scoring systems. That is, missing item responses are treated as incorrect in a sum score approach because a missing value does not add to the sum, ignored in a mean score approach because only observed responses are averaged, and shrunken towards the mean and weighted by difficulty in an Empirical Bayes IRT theta score, whereas the EIRM simply models the available item responses directly. The comparison of these four methods in the presence of missing data is intended to complement existing approaches for addressing missing item response data, rather than suggest that missing item response data can or should be ignored. For example, multiple imputation in IRT (Finch, 2008; Sulis & Porcu, 2017) and full information maximum likelihood (FIML; Enders & Bandalos, 2001) are powerful methods for addressing missing item response data but are not evaluated here to maintain focus on how CTT and IRT-based scoring systems perform in the presence of data that is missing completely at random (MCAR), which is theoretically ignorable, or missing at random (MAR), which is not (Holman & Glass, 2005), given prior research suggesting that IRT-based methods are more robust to missing data (de Bock, et al., 2016).

In short, this study seeks to answer an important methodological question with implications for the applied researcher. That is, under what conditions, if any, do the theoretical advantages of the EIRM obtain, and to what extent are they worth the additional interpretational complexity and required computational power when compared to the traditional two-step approach? If so, the EIRM provides a potentially powerful method in the analyst's toolkit, and if not, simpler methods can be employed without loss of information.

Methods for Analyzing Test Score Data

Two-Step Procedures. In a two-step procedure, the latent trait of interest is estimated for each person and then analyzed as the outcome variable using a standard statistical model such as OLS regression (Christensen, 2006, p. 185; Ye, 2016, p. 43). For example, consider the following regression model, in which Y_i represents an estimated latent trait score and β_1 represents an average treatment effect:

$$Y_i = \beta_0 + \beta_1 treat_i + \epsilon_i$$

$$\epsilon \sim (0, \sigma_e^2).$$

The estimated latent trait scores Y_i may be generated in a classical test theory (CTT) or item response theory (IRT) framework. In CTT, a sum (total correct) or mean score is employed, such that the observed score across all items X_j is equal to the sum of the responses $\Sigma_{j=1}^{J} X_j$ or the mean of the responses $\frac{1}{J}\Sigma_{j=1}^{J} X_j$. When there is no missing data, the sum and mean scores will be perfectly correlated. In IRT, the latent trait estimate, commonly denoted θ, is calculated by maximizing the likelihood of θ given the estimated item parameters (i.e., item difficulty, discrimination, and pseudo-guessing). Generally, the IRT approach has been argued to be superior because IRT theta estimates are on an interval scale (Jabrayliov, Emons, & Sijtsma, 2016; Ferrando & Chico, 2007; Harwell & Gatti, 2001), though empirically, differences between the CTT and IRT scoring are often found to be minor (Xu & Stone, 2012; Sebille, et al., 2010). One potential limitation of the two-step analysis approach is that, regardless of what type of scoring procedure is used to estimate the latent trait, the outcome variable is treated as known when it is measured with uncertainty (i.e., measurement error), and therefore, resulting estimates may be biased (Embretson, 1996).

The Explanatory Item Response Model (EIRM). The EIRM is a cross-classified multilevel logistic regression model, in which item responses are nested within the cross-classification of persons and items. In its simplest form with random effects for persons and items and without predictors, it can be expressed as

$$logit\left(P(y_{ij}=1)\right)=\beta_0+\theta_j+\zeta_i$$

$$\theta_j \sim N(0,\sigma_\theta^2)$$

$$\zeta_i \sim N(0,\sigma_\zeta^2)$$

in which the log-odds of a correct response to item i for person j is a function of a constant term β_0, person ability θ_j and item easiness ζ_i (item easiness is the negative of the more familiar item difficulty parameter in traditional IRT modelling). The EIRM with no person or item predictors is equivalent to the Rasch or One-Parameter Logistic (1PL) IRT model when the item easiness parameters are considered fixed.

Following the taxonomy of Wilson, De Boeck, and Carstensen (2008, p. 95), the EIRM with random person and item effects is called a “doubly descriptive” model, as it solely provides estimates of the variances of both persons and items without any variables to *explain* systematic differences in person ability or item easiness. The EIRM becomes “person explanatory” or “item explanatory” when predictors at the person or item level are added to the model, or “doubly explanatory” when both person and item level predictors are included. For example, an EIRM could be used to test whether older students have higher performance (person explanatory), whether algebra or geometry problems are more difficult (item explanatory), or both (doubly explanatory). For the purposes of this study, we can explore questions of causal inference by employing a person explanatory model that includes a person-level coefficient β_1 that estimates the causal effect of treatment on the log-odds of a correct response:

$$logit\left(P\left(y_{ij}=1\right)\right)=\beta_0+\beta_1 treat_i+\theta_j+\zeta_i$$

$$\theta_j \sim N\left(0,\sigma_\theta^2\right)$$

$$\zeta_i \sim N(0,\sigma_\zeta^2).$$

While more complex to interpret due to the cross-classified multilevel structure and logistic link function, and more computationally demanding than a CTT-based analysis such as sum or mean scores due to the numerical integration required in parameter estimation, the EIRM provides a key theoretical advantage in the analysis of test score data. That is, various studies have demonstrated that the EIRM deattenuates estimates of regression coefficients, thus counteracting the effects of measurement error compared to regression on observed scores (Zwinderman, 1991; Christensen, 2006; Briggs, 2008), suggesting that the EIRM may provide a more sensitive test of between-group differences such as causal treatment effects. Measurement error is an important issue in data analysis, yet strikingly, in some fields, it is barely addressed, with one systematic review reporting that only 7% of studies investigated or corrected for measurement error (Brakenhoff et al., 2018, in epidemiology). Similarly, while CTT-based scoring approaches such as sum scoring are commonly employed, they have long been known to have issues with “adverse effects on validity, reliability, and qualitative classification” (McNeish & Wolf, 2020, p. 2287). Thus, the EIRM may provide a straightforward way to gain more robust and fine-grained insight into intervention effects than other approaches, a hypothesis we now test through simulation and application to empirical assessment data.

Methods

Data Generating Process

The simulation and analytic procedures were implemented in R (R Core Team 2022). In total, we simulated 252,000 data sets and applied four analytic models—sum score, mean score, One Parameter Logistic (1PL) IRT theta score, and EIRM—to each, for a total of 1,008,000 results. We provide a detailed replication toolkit containing all R code necessary for researchers to replicate or extend the simulation or explore the empirical data application. We employed a full factorial design to assess the performance of each model under an array of conditions representing what a researcher might encounter including violations of model assumptions to probe the limits of the applicability and value of the EIRM under potential misspecification. The simulation factors include small, moderate, and large sample sizes (300, 500, 1000), short, moderate, and long assessment lengths (10, 20, 40 items), null and positive treatment effect sizes (0 or 0.2 standard deviations on the latent trait), a range of missing item response rates (0%, 5%, 10%, 25%), normal or skewed latent trait distributions, homoscedastic or heteroskedastic variances by group (in which the treatment group standard deviation is twice that of the control group), and MCAR or MAR missing data mechanisms (in which more difficult items have a higher probably of missingness). Missing item responses could represent either unanswered questions or items randomly drawn and administered from a larger test bank. The data-generating process was based a two-group treatment-control design in which items were equally correlated with the latent trait (i.e., a 1PL or Rasch model). The null hypothesis rejection rates provide false positive rates when the treatment effect is 0, and statistical power when the treatment effect is positive. IRT models were estimated with the `mirt` R package (Chalmers, 2012), and the EIRM with the `glmer` function from the `lme4` package (Bates, et al., 2014). For a tutorial on estimating the EIRM in R, see De Boeck et al. (2011).

Estimating Treatment Effects

For each simulated data set, we estimated the treatment effect and associated *t*-statistic, *p*-value, and whether the null hypothesis was rejected (at the 5% level, one sided) under four model parameterizations listed in Table 1, which includes reference R code for fitting each model.

[Insert Table 1 here]

In all models, the parameter of interest is the treatment effect β_1, and the errors (e_i, ζ_i, θ_j) are assumed to be normally distributed with mean 0 and constant variance and uncorrelated with the predictors (or other random effects in the case of the EIRM). The models for the sum score, mean score, and theta score are identical OLS regression models, and the EIRM is estimated as a generalized linear mixed model (GLMM) with a logit link, a random effect for student ability (θ_j) cross-classified with random effects for item easiness (ζ_i). Note that the EIRM is mathematically equivalent to a 1PL IRT model with random item difficulties (De Boeck, 2008), and that by including the item random effect (ζ_i), we best approximate the data-generating process, as the item difficulties were randomly drawn from a normal distribution. Across all models, we assess the statistical significance of the treatment effects via standard t-tests for the OLS regression models (sum, mean, and theta scores) and Wald tests for the EIRM.

Table 2 shows the treatment effect estimates for the four models fit to a single simulated data set with 100 subjects, 20 items, a treatment effect size of +0.20 standard deviations on the latent trait, a missing item response rate of 10% (MCAR), and a normally distributed and homoscedastic latent trait distribution. The coefficients on the treatment effects cannot be compared directly as they are all on different scales and cannot be converted to the scale of the underlying latent trait (Mood, 2009, pp. 73-74). However, because the data is identical and each model attempts to assess the same underlying treatment effect parameter, the *t*-statistics and *p*-

values can be compared directly. In this single simulation, we see that while the differences are small, the *t* is largest and *p* is smallest for the EIRM, suggesting it may provide more statistical power than the alternative approaches. An analysis of multiple simulations will shed light on the strength and consistency of this pattern and enable how estimates of bias, standard error calibration, false positive rates, and statistical power differ across methods and may depend on the various simulation factors.

[insert Table 2 here]

Results

The figures and tables below present the most salient characteristics of the four models in terms of (a) parameter bias, (b) standard error calibration, (c) false positive rates, and (d) statistical power. Detailed tables in the Online Supplemental Materials (OSM) provide additional descriptive statistics across additional combinations of the simulation factors for each of the performance metrics.

Bias

Because there is no analytic method to convert the model-based treatment effect point estimates to the scale of the latent trait underlying the item responses (Mood, 2009, pp. 73-74), we can most easily analyze potential bias in the treatment effect coefficient by analyzing the simulations in which the true treatment effect is 0, because, while each model is on a different scale, the parameter estimates will be proportional to true value of 0 for all models (ibid). Each estimate was divided by the standard error times the square root of the sample size so that the magnitude of bias could be compared across methods in standard deviation units.

Figure 1 shows the estimated bias by method according faceted by skewness and heteroskedasticity. Overall, bias appears comparable across all methods. When the group variances are homoscedastic, the estimated bias is almost zero, but when the latent trait is both skewed and heteroskedastic, there is a downward bias of approximately 0.07 standard deviations for all methods.

[Insert Figure 1 here]

Standard Error (SE) Calibration

To assess the calibration of the SEs of each method, we compare the mean model-based (asymptotic) SE to the true SE (i.e., the observed standard deviation of the point estimates) for each condition. The model-based SE is expressed as a percentage of the true SE, so that a value of 100% would indicate that the model-based SE is identical to the true SE. Figure 2 shows the distribution of SE calibration for each method again faceted by skewness and heteroskedasticity. We see that under homoscedasticity, the model-based SEs are close to their true values for all methods, within about 10 percentage points. Heteroskedasticity results in underestimated SEs for all models, though the effects are less pronounced for the EIRM and the theta scores than the sum and mean scores. This result suggests that in the presence of heteroskedasticity, either robust methods (e.g., the `estimatr` R package developed by Blair, et al., 2019) or nonparametric resampling approaches such as bootstrapping should be employed to obtain accurate SEs.

[Insert Figure 2 here]

False Positive Rates

Figure 3 shows the average false positive rates for each method again faceted by skewness and heteroskedasticity. Following directly from the results of the bias and SE calibration figures above, under homoskedasticity, the false positive rates are all extremely close to the nominal value of 0.05. In the bottom left quadrant (heteroskedasticity and normality), the false positive rates are slightly inflated for the EIRM, theta, and mean scores, and significantly inflated for the sum score. The bottom right quadrant (heteroskedasticity and skewness) shows that all methods provide extremely low false positive rates, but this is a function of the downward bias on the treatment effect point estimate under these conditions observed in Figure 1.

[Insert Figure 3 here]

Statistical Power

The downwardly biased SEs produced by model misspecification would result in inflated estimates of statistical power if taken at face value. Thus, we used the true SEs (i.e., the standard deviation of the point estimates for each condition) to recalculate the z-statistics and associated hypothesis tests for each simulated model, akin to a resampling approach for estimating uncertainty that is robust to model misspecification. Figure 4 shows the statistical power for each method as a function of the missing item response rate, faceted by skewness, heteroskedasticity, and whether the items were missing at random (MAR).

The absolute and relative statistical power provided by each model varies substantially across these eight conditions. For example, in the topmost left plot, when all model assumptions are met, the differences between each method are small, but the EIRM and the theta score provide slightly more power than the sum and mean scores providing a benefit of about one percentage point with complete data. When heteroskedasticity is present, but there is no skew and missing data are MCAR, the power advantage of the EIRM and theta scores is substantial, at about 10 percentage points. When both skewness and heteroskedasticity are present, the downward bias on the treatment effect point estimate observed in Figure 1 reduces power for all methods to near floor levels. Furthermore, the performance of the sum and mean scores diverge depending on whether the data is MAR or MCAR, with the sum score providing more power under MAR. In sum, there appears to be no single model that performs best across all conditions, but rather, the performance of the models is highly dependent upon the degree and kind of model misspecification. As such, exploratory data analysis that checks for skewness, heteroskedasticity, and MAR data should be conducted prior to model selection and the determination whether to use model-based or robust standard errors.

[insert Figure 4 here]

Regression Models

To better understand the divergent patterns of statistical power presented in Figure 4, we next fit a series of regression models to examine the precise size and statistical significance of statistical power differences each method under the eight conditions of model misspecification (heteroskedasticity crossed with skew crossed with MAR). For ease of interpretation, we fit separate regression models to each of the eight misspecification conditions rather than include three- and four-way interactions in a single model. We used the following multilevel linear probability model,

$$reject_{ij} = \beta_0 + \beta_1 method_{ij} + \beta_2 miss\_rate_j + \beta_3 method \times miss\_rate_{ij} +$$

$$\beta_4 n\_subjects_j + \beta_5 n\_items_j + \zeta_j + \epsilon_{ij}$$

in which $reject_{ij}$ is a 0/1 variable (transformed to 0/100 so coefficients can be interpreted in percentage points) indicating whether the null hypothesis was rejected for model i in data set j. Each β represents the main effect of the variable on power, β_3 captures the interaction between estimation method and missing item response rate, ζ_j is a random effect for each data set to account for the fact that multiple models were fit to the same simulated data set, and ϵ_{ij} is the residual error. Due to the bounded nature of the outcome and the non-linearity inherent in power analysis (i.e., power has an upper asymptote of 1), we treated all predictor variables as polytomous categories rather than continuous. We employed a linear probability model rather than a logistic or probit model for ease of interpretation of the coefficients. (An alternative probit model specification with transformed continuous sample size as a predictor is included in the OSM as a sensitivity check; all substantive findings are identical as the direction and relative magnitude of relevant effects is unchanged). Importantly, because the data-generating process is known, the regression coefficients have a causal interpretation.

The full output of the fitted models is presented in Table 3 and demonstrate several insights. First, in the baseline (i.e., appropriately specified) model, with no missing data, the EIRM provides a modest but statistically significant benefit to power over the sum and mean scores of approximately one percentage point and is not significantly different from the theta score. Missing item response rates appear most deleterious to the power of the sum score method, reducing power by approximately 2.5 percentage points more than the EIRM when the missing item response rate is 25%. When heteroskedasticity is present, the power advantage of the EIRM and the theta score over the sum and mean scores is nearly ten percentage points with complete data. Interestingly, when the latent trait is skewed, the mean and sum scores provide more power than the EIRM or theta score, but again the difference is small. The models that include both heteroskedasticity and skew have near floor levels of power due to the negative bias encountered earlier. When missing item response data is MAR, the mean score suffers much more than the sum score, which is expected in this case because missingness is predicted by ability, so the sum score's implicit treatment of a missing response as "incorrect" provides a benefit, but this finding would not generalize to other mechanisms of missingness. In sum, there is no single method that provides the "best" approach because model performance is dependent on the degree of misspecification, a finding that again underscores the need for appropriate exploratory data analysis.

[insert Table 3 here]

Application to Empirical Student Assessment Data

The results of the simulation show that the performance of each method is dependent upon the conditions of the data generating process, with only slight benefits to statistical power for the EIRM under appropriate model specification. We conclude the exploration of these methods with an empirical data example to determine the extent to which the results of the simulation are consistent with a real-world application. We employ a public use file from Kim et al. (2022) that explores the causal effect of the Model of Reading Engagement (MORE) literacy intervention on 2nd grade students' content comprehension test scores on a researcher-designed reading comprehension assessment. The assessment included three reading passages followed by a total of 20 multiple choice items, and the study employed a cluster-randomized design with 30 schools and 2174 students. The authors assess the intention-to-treat (ITT) effect of the MORE intervention by fitting a multilevel model in which the outcome was the standardized sum score. The authors then employed an EIRM to test for differential treatment effects depending on the reading passage by estimating treatment by passage interaction effects. However, they did not use an EIRM to assess the overall treatment effect.

The MORE assessment data differs in several key respects from the setup of our simulations. First, as a cluster-randomized trial, the data have a hierarchical structure with treatment assigned at the school level. Second, there is no missing item response data. Third, they included a rich set of demographic covariates such as pretest scores, race, gender, SES, and other student characteristics to increase the precision of their ITT estimates.

In our reanalysis of the MORE study data, we contrast three simplified versions of the models fit in the original paper by modeling the outcome as a function of the treatment indicator, a reading pretest score, and relevant random effects. The model for the sum score and theta score is

$$y_{ij} = \beta_0 + \beta_1 treat_j + \beta_2 pretest_{ij} + \zeta_j + \epsilon_{ij}$$

in which y_{ij} is the sum or theta score for student i in school j , β_1 is the treatment effect, β_2 is the pretest reading score, ζ_j is the random effect for school j , and ϵ_{ij} is the student-level error. (We did not fit a model for the mean score because the mean and sum are the same in this case because there is no missing item response data.) The EIRM is modeled as

$$logit\left(P(y_{ijk} = 1)\right) = \beta_0 + \beta_1 treat_j + \beta_2 pretest_{ij} + \theta_{ij} + \zeta_j + \nu_k$$

in which y_{ijk} is the dichotomous item response for student i in school j to item k, the other parameters are analogous to those of the sum score model, θ_{ij} represents random student ability cross-classified with ν_k, random item easiness.

Given the dependence of the robustness of each method to the data-generating conditions observed in the simulation, we begin with exploratory data analysis to determine to what extent the standard parametric assumptions are plausibly met by this data. There is no missing data, so the distinction between MCAR vs. MAR is not relevant. In terms of heteroskedasticity, the standard deviations of the estimated theta scores for each group are very similar, at 0.79 for control students and 0.88 for treatment students, suggesting a nearly homoscedastic error distribution. Last, examination of density plots (not pictured) reveals approximately symmetrical, unimodal distributions of the estimated theta scores. Thus, the application of the EIRM should provide a moderately more powerful and precise estimate of the treatment effect based on the pattern of results observed in the simulation.

We fit each model and present the results in Table 4. In accordance with the results of the simulations under appropriate model specification, we see that while all models reject the null hypothesis of zero treatment effect, the EIRM provides the greater *t*-statistic and lower *p*-value (two-sided). These results are consistent with the benefits of the EIRM in assessing treatment effects in empirical data. That is, when item responses are available and model assumptions are met, they can be used to generate a modestly more powerful estimate of treatment impact.

[insert Table 4 here]

Discussion

The results of the simulation study paint a nuanced picture of the tradeoffs of different modeling choices. That is, there is no one model that performs best across all metrics in all circumstances. Rather, performance is dependent on the extent to which parametric assumptions are tenable. That is, when the assumptions of normality and homoscedasticity are met, the EIRM provides a modest benefit over sum and mean scores and a negligible advantage over theta scores in terms of statistical power. However, this modest benefit comes at the dual costs of computational power and interpretational difficulty of the EIRM, and thus the theta score may be a better choice for practitioners as it appears to capture most of the benefits of the EIRM but may be easier for most practitioners and audiences to interpret.

Heteroskedasticity emerged as the most consequential type of model misspecification in the simulation, as it induces a downward bias for SEs across all methods, but these effects are less severe for the EIRM and the theta scores. As such, in the presence of heteroskedasticity, robust or nonparametric methods such as resampling should be employed to obtain accurate estimates of uncertainty. Furthermore, when heteroskedasticity and skewness occur simultaneously, the downward bias in the treatment effect point estimates across all methods makes inference a challenge generally, suggesting the utility of more flexible models that require fewer assumptions.

Limitations and Future Directions

With the benefits and drawbacks of the EIRM across a wide variety of conditions now clear, there are many areas for potential expansion to other research contexts. For example, recent work has explored the potential value of the 2PL EIRM with random item discriminations (Petscher et al. 2020, in Mplus; Burkner, 2019, using R's `brms`) and estimating the 2PL EIRM with fixed item discriminations in R in a GLMM framework through the `mixedmirt` (Chalmers 2012) and `PLmixed` (Rockwood & Jeon 2019) packages, but such models are not widely used (see Huang, 2021 for the only published example) and may suffer from accuracy issues (Zhang, Ackerman, & Wang 2021). Another line of inquiry could include the consequences of heterogeneous treatment effects, such as differential treatment impact on different items or types of items. Such effects would be hidden in classical test theory sum or mean scores, but they could be modeled explicitly with the EIRM via person-item interactions or a treatment effect that varies randomly at the item level (Gilbert, Kim, & Miratrix, 2022). Furthermore, a comparison of the performance of the EIRM to that of other methods that combine measurement and analysis into a single step, such as structural equation models (SEM) with latent variables could provide further insight into the relative strengths and weaknesses of such approaches. Unfortunately, at this time, the primary R package for estimating SEMs, `lavaan` (Rosseel 2012), does not support logistic models for binary indicators that would be most comparable to the EIRM approach.

A final challenge of the application of the EIRM involves the interpretation of the coefficients of the fitted models, as treatment effect coefficients on the logit scale may be more difficult to explain and justify to practitioners than the more familiar sum or mean score. As such, we suggest the following two approaches to make the EIRM results more interpretable. First, the fitted models can be used to estimate population-averaged response probabilities (e.g., using the `ggeffects` R package described in Lüdecke, 2018), representing treatment-control differences on the probability scale. Second, EIRM treatment effect coefficients can be converted to a Cohen's *d* type effect size by the process of "y-standardization" (see Breen, Karlson, & Holm, 2018 for the single-level case; see Hox, Moerbeek, & Van de Schoot, 2017, Chapter 6 for the multilevel case), whereby the logit-scale coefficient β_{logit} is divided by the estimated total standard deviation of a latent continuous variable Y* that could give rise to the observed dichotomous response Y, using the following formula

$$\beta_{ystd} = \frac{\beta_{logit}}{SD(Y^*)} = \frac{\beta_{logit}}{\sqrt{\frac{\pi^2}{3} + \sigma_\theta^2 + \sigma_{\zeta_0}^2 + \sigma_F^2}}$$

in which $\frac{\pi^2}{3} = 3.29$ is the variance of the logistic distribution, the σ_θ^2 and $\sigma_{\zeta_0}^2$ represent the variance components of the persons and items, and σ_F^2 is the variance of the fixed effects (i.e., the variance of the estimated linear predictor on the logit scale).

Conclusion

The Explanatory Item Response Model provides a potentially useful tool for the applied researcher in estimating treatment effects, but its applicability may be tempered by the tradeoffs involved in its use in diverse contexts. The EIRM can provide a modest benefit to statistical power when parametric assumptions are met and missing data is present, and a significant benefit to power under heteroskedasticity, but these benefits come at the cost of increased computational intensity and interpretational difficulty. Crucially, the performance of the EIRM typically only marginally superior to that of the IRT theta score, which appears to provide most of the benefits and is much more straightforward to interpret. When the models are misspecified, are not met, the EIRM and theta scores do provide more robust estimates of uncertainty than sum or mean scores, but more accurate estimates of uncertainty can also be achieved through other means such as robust or nonparametric methods. In sum, when only person-level variables are of interest, the benefits of the EIRM appear to be minor and demand a careful consideration of the tenability of model assumptions for appropriate use.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Blair, G., Cooper, J., Coppock, A., Humphreys, M., & Sonnet, L. (2019). Estimatr: Fast estimators for design-based inference. *R package version*.

Brakenhoff, T. B., Mitroiu, M., Keogh, R. H., Moons, K. G., Groenwold, R. H., & van Smeden, M. (2018). Measurement error is often neglected in medical literature: a systematic review. *Journal of clinical epidemiology*, *98*, 89-97.

Breen, R., Karlson, K. B., & Holm, A. (2018). Interpreting and understanding logits, probits, and other nonlinear probability models. *Annual Review of Sociology*, *44*, 39-54.

Briggs, D. C. 2008. "Using Explanatory Item Response Models to Analyze Group Differences in Science Achievement." *Applied Measurement in Education* 21 (2): 89–118. https://doi.org/10.1080/08957340801926086.

Bürkner, P. C. (2019). Bayesian item response modeling in R with brms and Stan. *arXiv preprint arXiv:1905.09501*.

Chalmers, R. P. 2012. "**Mirt**: A Multidimensional Item Response Theory Package for the*R*Environment." *Journal of Statistical Software* 48 (6). https://doi.org/10.18637/jss.v048.i06.

Christensen, K. B. (2006). From Rasch scores to regression. *Journal of Applied Measurement*, *7*(2), 184.

de Bock, E., Hardouin, J. B., Blanchin, M., Le Neel, T., Kubis, G., Bonnaud-Antignac, A., ... & Sebille, V. (2016). Rasch-family models are more valuable than score-based approaches for analysing longitudinal patient-reported outcomes with missing data. *Statistical methods in medical research*, *25*(5), 2067-2087.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*, 1-28.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533-559.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial anova designs. *Applied Psychological Measurement, 20(3)*, 201-212.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural equation modeling*, *8*(3), 430-457.

Ferrando, P. J., & Chico, E. (2007). The External Validity of Scores Based on the Two-Parameter Logistic Model: Some Comparisons between IRT and CTT. *Psicologica: International Journal of Methodology and Experimental Psychology*, *28*(2), 237-257.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225-245.

Gilbert, J. B., Kim, J. S., and Miratrix, L. W. (2022). Modeling Item-Level Heterogeneous Treatment Effects with the Explanatory Item Response Model: Leveraging Online Formative Assessments to Pinpoint the Impact of Educational Interventions. (EdWorkingPaper: 22-619). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/m3jh-kh96

Hartig, J., Frey, A., Nold, G., & Klieme, E. 2011. “An Application of Explanatory Item Response Modeling for Model-Based Proficiency Scaling.” *Educational and Psychological Measurement* 72 (4): 665–86. https://doi.org/10.1177/0013164411430707.

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, *71*, 105-131.

Holman, R., & Glas, C. A. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 1-17.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Huang, S. (2021). *Modern Applications of Cross-classified Multilevel Models (CCMMs) in Social and Behavioral Research: Illustrations with R Package PLmixed*. University of California, Los Angeles.

Jabrayilov, R., Emons, W. H., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied psychological measurement*, *40*(8), 559-572.

Kim, J. S., Burkhauser, Burkhauser, M. A., Relyea, J. E., Gilbert, J. B., Scherer, E., Fitzgerald, J., Mosher, D., & McIntyre, J. (2022). A longitudinal randomized trial of a sustained content literacy intervention from first to second grade: Transfer effects on students’ reading comprehension. *Journal of Educational Psychology*. Advance Online Publication.

Kim, J. S., Burkhauser, M. A., Relyea, J. E., Gilbert, J. B., Scherer, E., Fitzgerald, J., Mosher, D., & McIntyre, J. (2022). “Replication Data for: A Longitudinal Randomized Trial of a

Sustained Content Literacy Intervention from First to Second Grade: Transfer Effects on Students' Reading Comprehension Outcomes." Harvard Dataverse. https://doi.org/10.7910/DVN/LAWFFU.

Lüdecke, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, *3*(26), 772.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior research methods*, *52*(6), 2287-2305.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European sociological review*, *26*(1), 67-82.

Petscher, Y., Compton, D. L., Steacy, L., & Kinnon, H. 2020. "Past Perspectives and New Opportunities for the Explanatory Item Response Model." *Annals of Dyslexia* 70 (2): 160–79. https://doi.org/10.1007/s11881-020-00204-y.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rabbitt, M. P. (2018). Causal inference with latent variables from the Rasch model as outcomes. *Measurement*, *120*, 193-205.

Randall, J., Fai Cheong, Y., & Engelhard, G. 2011. "Using Explanatory Item Response Theory Modeling to Investigate Context Effects of Differential Item Functioning for Students With Disabilities." *Educational and Psychological Measurement* 71 (1): 129–47. https://doi.org/10.1177/0013164410391577.

Rockwood, N. J., & Jeon, M. 2019. "Estimating Complex Measurement and Growth Models Using the R Package PLmixed." *Multivariate Behavioral Research* 54 (2): 288–306. https://doi.org/10.1080/00273171.2018.1516541.

Rosseel, Y. 2012. “Lavaan: An r Package for Structural Equation Modeling” 48. https://doi.org/10.18637/jss.v048.i02.

Sébille, V., Hardouin, J. B., Le Néel, T., Kubis, G., Boyer, F., Guillemin, F., & Falissard, B. (2010). Methodological issues regarding power of classical test theory (CTT) and item response theory (IRT)-based approaches for the comparison of patient-reported outcomes in two groups of patients-a simulation study. *BMC medical research methodology*, *10*(1), 1-10.

Stevenson, C. E., Hickendorff, M., Resing, W. C., Heiser, W. J., & de Boeck, P. A. (2013). Explanatory item response modeling of children's change on a dynamic test of analogical reasoning. *Intelligence*, *41*(3), 157-168.

Sulis, I., & Porcu, M. (2017). Handling missing data in item response theory. Assessing the accuracy of a multiple imputation procedure based on latent class analysis. *Journal of Classification*, *34*(2), 327-359.

Wilson, M., & De Boeck, P. 2004. “Descriptive and Explanatory Item Response Models.” In, 43–74. Springer New York. https://doi.org/10.1007/978-1-4757-3990-9_2.

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. *Assessment of competencies in educational contexts*, 91-120.

Xu, T., & Stone, C. A. (2012). Using IRT trait estimates versus summated scores in predicting outcomes. *Educational and Psychological Measurement*, *72*(3), 453-468.

Ye, F. (2016). Latent growth curve analysis with dichotomous items: Comparing four approaches. *British Journal of Mathematical and Statistical Psychology*, *69*(1), 43-61.

Zhang, J., Ackerman, T., & Wang, Y. 2021. “2PL Model: Compare Generalized Linear Mixed Model with Latent Variable Model Based IRT Framework.” http://dx.doi.org/10.31234/osf.io/p6wuz.

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*(4), 589-600.

Table 1

Statistical models and R code for the four models evaluated in this study: sum score, mean score, Item Response Theory theta score, and Explanatory Item Response Model (EIRM)

Analytic Method	Statistical Model	Sample R Code
Sum Score	$sum_i = \beta_0 + \beta_1 treat_i + e_i$	`lm(sum ~ treat, data)`
Mean Score	$mean_i = \beta_0 + \beta_1 treat_i + e_i$	`lm(mean ~ treat, data)`
Theta Score	$\theta_i = \beta_0 + \beta_1 treat_i + e_i$	`lm(theta ~ treat, data)`

EIRM	$logit\left(P\left(Y_{ij} = 1\right)\right)$
	$= \beta_0 + \beta_1 treat_i + \zeta_i$
	$+ \theta_j$
		`glmer(response ~ treat +`
		`(1|person_id) + (1|item_id),`
		`data, family = “binomial”)`

Table 2

Comparison of treatment effect estimates based on four analytic models: sum score, mean score, IRT theta score, and explanatory item response model (EIRM) fit to a single simulated data set

Method	Estimate	SE	t-statistic	p-value
Sum Score	0.592	0.221	2.673	0.008
Mean Score	0.063	0.024	2.578	0.010
Theta Score	0.391	0.137	2.858	0.004
EIRM	0.501	0.172	2.913	0.004

Note. Simulation included 500 Subjects, 10 Items, +0.20 Treatment Effect on the Latent Trait, 10% MCAR item responses. SE = standard error; MCAR = Missing Completely at Random.

Table 3

Multilevel linear probability model statistical power by analytic method

	Baseline		**Het.**		**Skew**		**Het. & Skew**		**MAR**		**Het. & MAR**		**Skew & MAR**		**Het. & Skew & MAR**	
Predictors	*Est.*	*SE*	*Est.*	*SE*	*Est.*	*SE*	*Est.*	*SE*	*Est.*	*SE*	*Est.*	*SE*	*Est.*	*SE*	*Est.*	*SE*
Intercept	45.83 ***	0.90	22.87 ***	0.92	54.61 ***	0.83	5.35 ***	0.40	45.41 ***	0.98	23.59 ***	0.98	51.14 ***	0.89	5.51 ***	0.46
1 = Mean Score	-0.98 **	0.31	-9.49 ***	0.44	1.09 ***	0.31	-2.42 ***	0.22	-1.11 ***	0.31	-14.02 ***	0.51	1.07 ***	0.32	-3.27 ***	0.27
1 = Sum Score	-0.98 **	0.31	-9.49 ***	0.44	1.13 ***	0.31	-2.42 ***	0.22	-0.78 *	0.31	-6.67 ***	0.51	0.58	0.32	-1.71 ***	0.27
1 = Theta Score	-0.29	0.31	-0.04	0.44	-0.53	0.31	-0.93 ***	0.22	-0.16	0.31	-0.38	0.51	-0.51	0.32	-1.16 ***	0.27
1 = 5% Missing	-1.18	0.92	-0.60	0.96	-1.56	0.85	-0.67	0.42								
1 = 10% Missing	0.82	0.92	-1.20	0.96	-2.58 **	0.85	-0.49	0.42	-1.27	0.93	2.56 **	0.97	0.27	0.85	0.11	0.46
1 = 25% Missing	-1.67	0.92	-2.87 **	0.96	-3.76 ***	0.85	-1.22 **	0.42	-3.58 ***	0.93	6.62 ***	0.97	-1.93 *	0.85	0.49	0.46
1 = 500 Subjects	16.30 ***	0.76	8.61 ***	0.76	18.48 ***	0.70	0.17	0.33	17.30 ***	0.89	9.02 ***	0.86	18.61 ***	0.81	-0.16	0.39

1 = 1000 Subjects	40.04 ***	0.76	21.86 ***	0.76	37.50 ***	0.70	0.15	0.33	40.99 ***	0.89	23.47 ***	0.86	38.61 ***	0.81	-0.02	0.39
1 = 20 Items	5.17 ***	0.76	4.63 ***	0.76	4.78 ***	0.70	0.14	0.33	3.87 ***	0.89	4.54 ***	0.86	4.87 ***	0.81	-0.06	0.39
1 = 40 Items	7.54 ***	0.76	8.35 ***	0.76	5.88 ***	0.70	1.57 ***	0.33	7.61 ***	0.89	8.46 ***	0.86	8.42 ***	0.81	1.49 ***	0.39
Mean x 5% Missing	-0.04	0.44	0.58	0.62	-0.07	0.44	0.44	0.31								
Sum x 5% Missing	0.16	0.44	0.60	0.62	-1.09 *	0.44	0.47	0.31								
Theta x 5% Missing	0.13	0.44	0.09	0.62	0.09	0.44	-0.02	0.31								
Mean x 10% Missing	-0.11	0.44	1.53 *	0.62	-0.29	0.44	0.31	0.31	-0.71	0.44	-2.60 ***	0.73	0.58	0.45	-0.22	0.39
Sum x 10% Missing	-0.91 *	0.44	1.02	0.62	-1.11 *	0.44	0.20	0.31	0.13	0.44	4.47 ***	0.73	0.18	0.45	1.13 **	0.39
Theta x 10% Missing	-0.04	0.44	0.04	0.62	0.04	0.44	0.13	0.31	-0.09	0.44	0.20	0.73	-0.09	0.45	0.11	0.39
Mean x 25% Missing	-0.27	0.44	0.64	0.62	-1.07 *	0.44	0.96 **	0.31	-1.67 ***	0.44	-16.80 ***	0.73	2.42 ***	0.45	-1.69 ***	0.39
Sum x 25% Missing	-2.64 ***	0.44	0.33	0.62	-3.02 ***	0.44	1.02 **	0.31	0.67	0.44	13.07 ***	0.73	-1.11 *	0.45	5.64 ***	0.39
Theta x 25% Missing	-0.13	0.44	0.09	0.62	-0.24	0.44	0.24	0.31	-0.24	0.44	0.76	0.73	-0.16	0.45	0.47	0.39

Random Effects

σ^2	215.20	430.06	220.81	109.55	215.34	594.02	232.06	168.23
τ_{00}	1695.92 $_{id}$	1623.97 $_{id}$	1418.03 $_{id}$	293.32 $_{id}$	1722.01 $_{id}$	1515.50 $_{id}$	1404.71 $_{id}$	306.46 $_{id}$
ICC	0.89	0.79	0.87	0.73	0.89	0.72	0.86	0.65
N	18000 $_{id}$	18000 $_{id}$	18000 $_{id}$	18000 $_{id}$	13500 $_{id}$	13500 $_{id}$	13500 $_{id}$	13500 $_{id}$
Observations	72000	72000	72000	72000	54000	54000	54000	54000

* $p<0.05$ ** $p<0.01$ *** $p<0.001$

Note: Reference groups are EIRM, 0% missing data, and 10 test items. Baseline = appropriately specified model. Het = heteroskedasticity; MAR = missing at random.

Table 4

Comparison of treatment effect estimates based on sum score, theta score, and explanatory item response model (EIRM) analytic methods for the Model of Reading Engagement (MORE) literacy intervention data.

Method	Estimate	SE	t-statistic	p-value
Sum Score	0.154	0.069	2.219	0.031
Theta Score	0.135	0.058	2.332	0.024
EIRM	0.176	0.074	2.397	0.017

Note. SE = standard error. Mean scores were not included in this table because there were no missing item responses and thus are equivalent to the sum scores in this case.

Figure 1

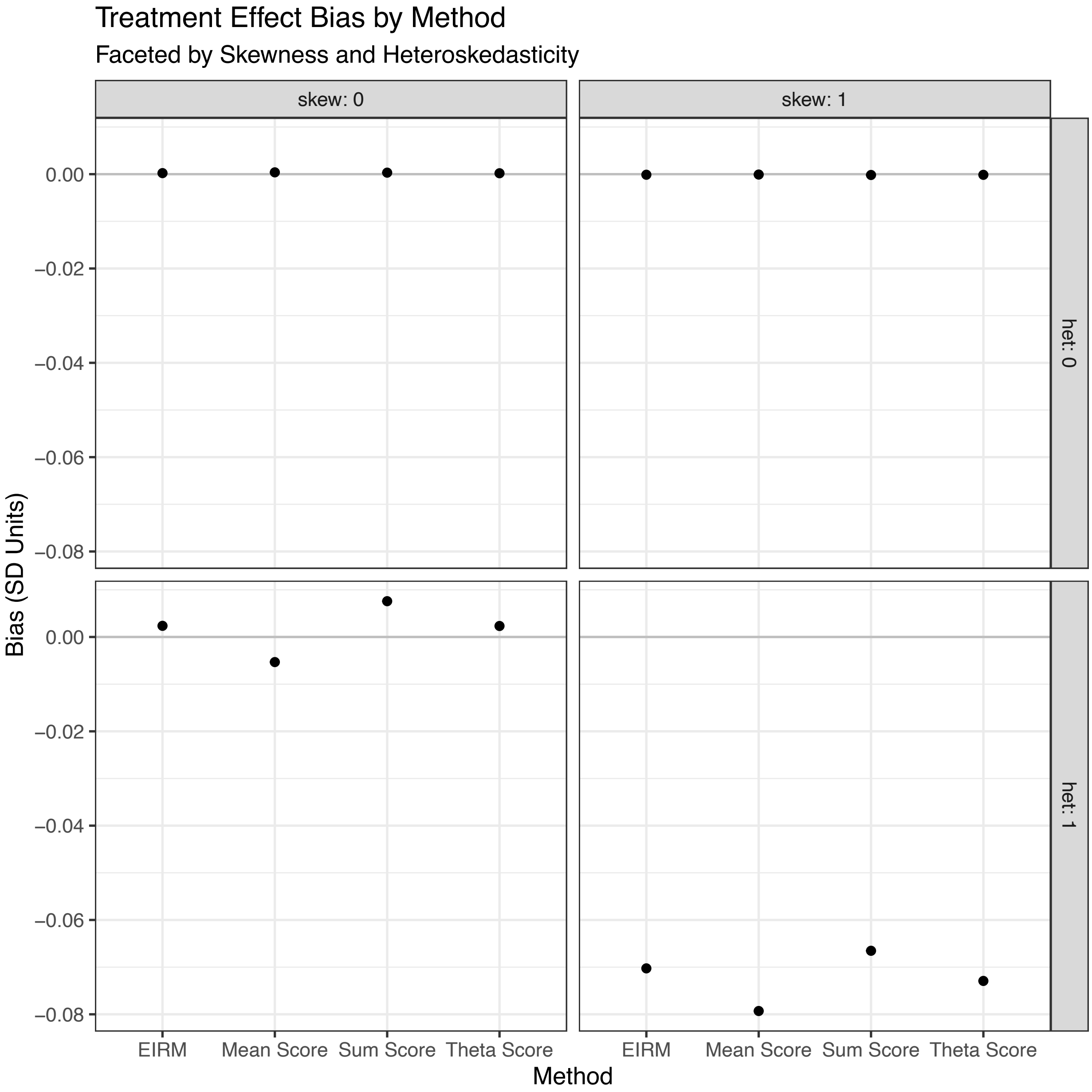

Treatment Effect Bias by Method
Faceted by Skewness and Heteroskedasticity
skew: 0
skew: 1
het: 0
het: 1
Bias (SD Units)
0.00
−0.02
−0.04
−0.06
−0.08
EIRM
Mean Score
Sum Score
Theta Score
Method

Figure 2

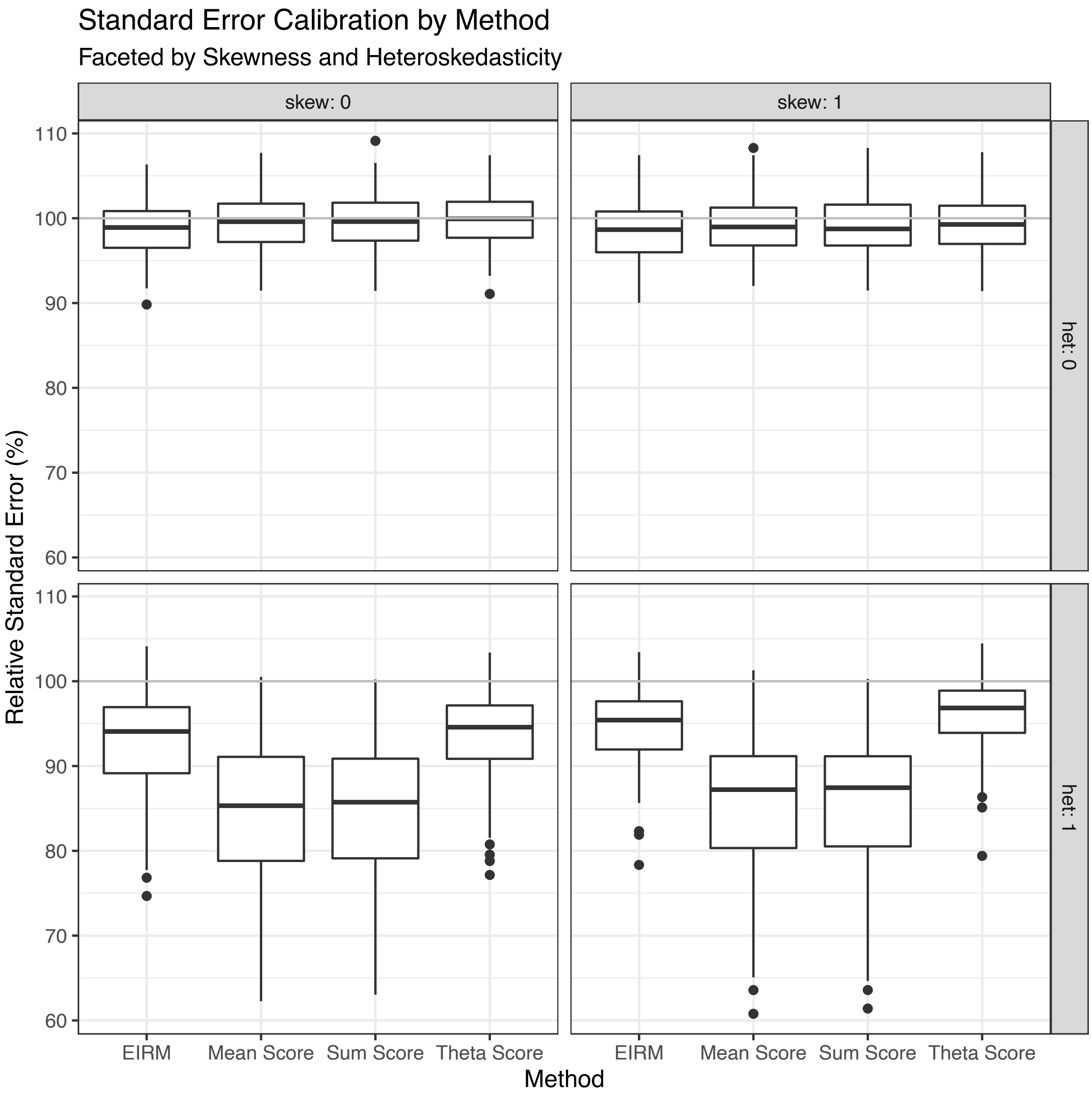

Standard Error Calibration by Method
Faceted by Skewness and Heteroskedasticity
skew: 0
skew: 1
het: 0
het: 1
Relative Standard Error (%)
110
100
90
80
70
60
EIRM
Mean Score
Sum Score
Theta Score
Method

Figure 3

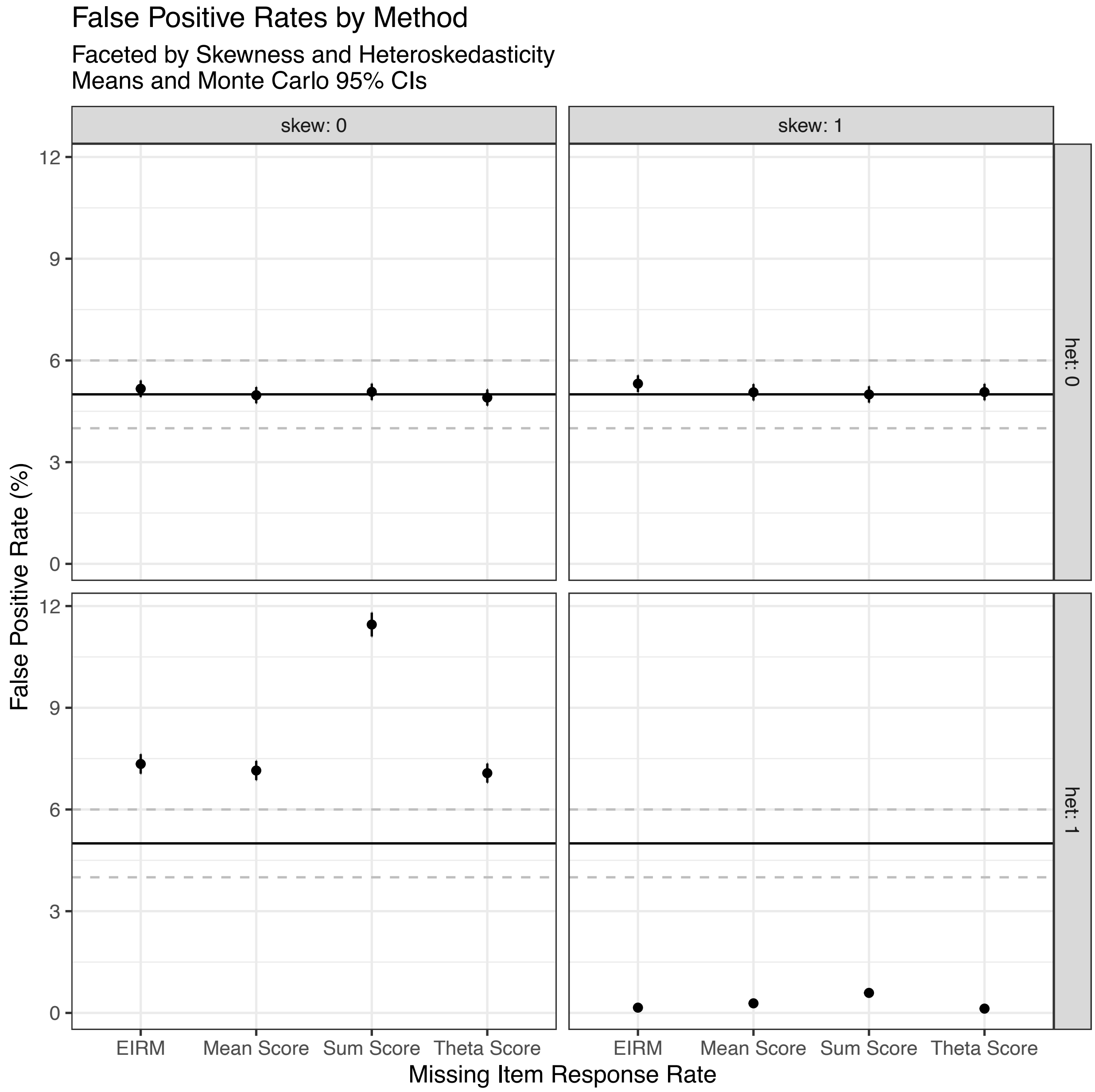
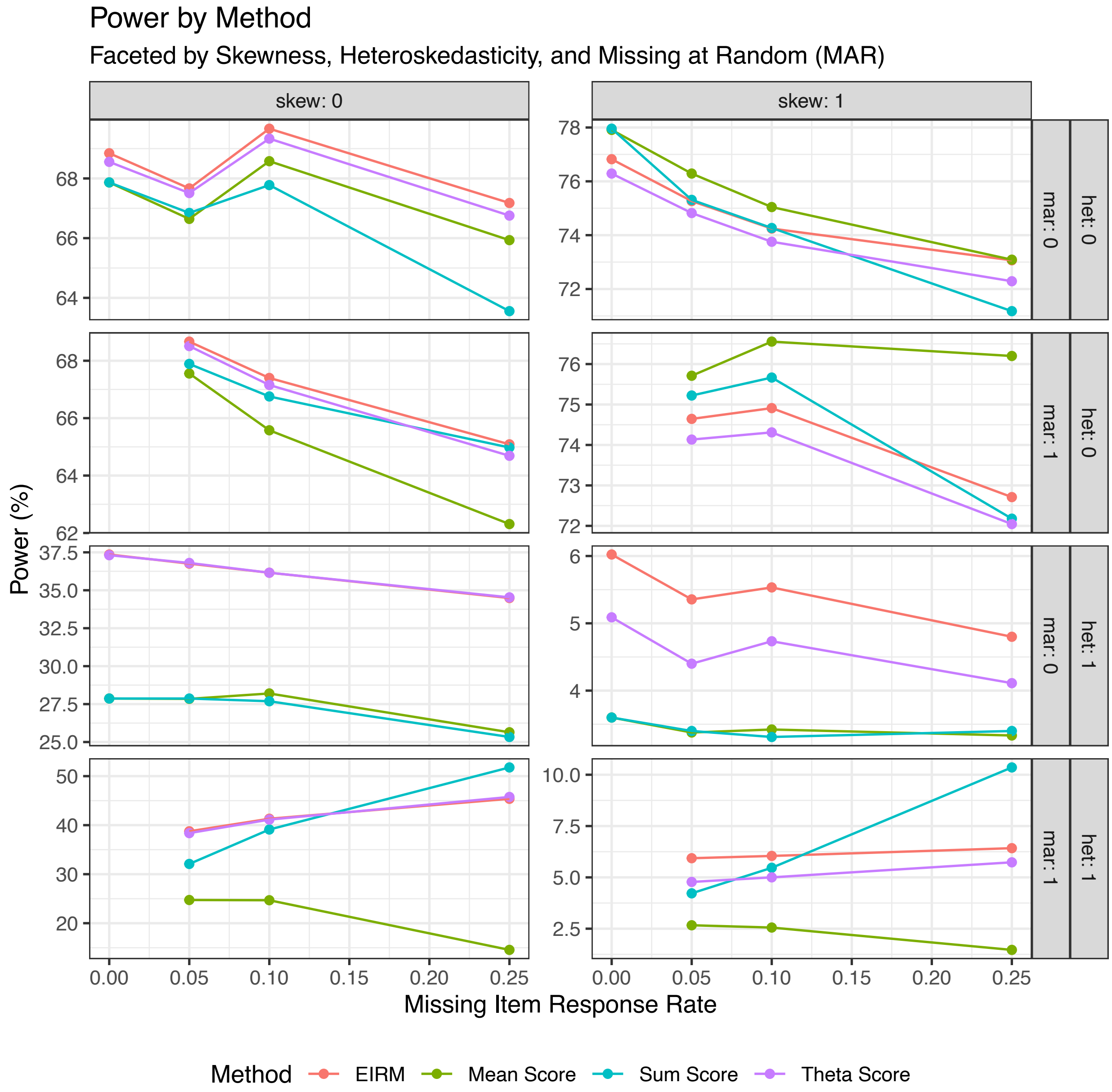

False Positive Rates by Method
Faceted by Skewness and Heteroskedasticity
Means and Monte Carlo 95% CIs
skew: 0
skew: 1
het: 0
het: 1
False Positive Rate (%)
0
3
6
9
12
EIRM
Mean Score
Sum Score
Theta Score
Missing Item Response Rate

Figure 4


Power by Method
Faceted by Skewness, Heteroskedasticity, and Missing at Random (MAR)
skew: 0
skew: 1
mar: 0
het: 0
mar: 1
het: 0
mar: 0
het: 1
mar: 1
het: 1
Power (%)
Missing Item Response Rate
0.00
0.05
0.10
0.15
0.20
0.25
64
66
68
62
37.5
35.0
32.5
30.0
27.5
25.0
50
40
30
20
78
76
74
72
75
73
6
5
4
10.0
7.5
5.0
2.5
Method
EIRM
Mean Score
Sum Score
Theta Score