

Psychological Methods

How Survey Scoring Decisions Can Influence Your Study's Results: A Trip Through the IRT Looking Glass

James Soland, Megan Kuhfeld, and Kelly Edwards

Online First Publication, July 14, 2022. <http://dx.doi.org/10.1037/met0000506>

CITATION

Soland, J., Kuhfeld, M., & Edwards, K. (2022, July 14). How Survey Scoring Decisions Can Influence Your Study's Results: A Trip Through the IRT Looking Glass. *Psychological Methods*. Advance online publication. <http://dx.doi.org/10.1037/met0000506>

How Survey Scoring Decisions Can Influence Your Study's Results: A Trip Through the IRT Looking Glass

James Soland^{1, 2}, Megan Kuhfeld², and Kelly Edwards¹

¹ School of Education and Human Development (EHD), University of Virginia

² Collaborative for Student Growth, NWEA Portland, Oregon, United States

Abstract

Though much effort is often put into designing psychological studies, the measurement model and scoring approach employed are often an afterthought, especially when short survey scales are used (Flake & Fried, 2020). One possible reason that measurement gets downplayed is that there is generally little understanding of how calibration/scoring approaches could impact common estimands of interest, including treatment effect estimates, beyond random noise due to measurement error. Another possible reason is that the process of scoring is complicated, involving selecting a suitable measurement model, calibrating its parameters, then deciding how to generate a score, all steps that occur before the score is even used to examine the desired psychological phenomenon. In this study, we provide three motivating examples where surveys are used to understand individuals' underlying social emotional and/or personality constructs to demonstrate the potential consequences of measurement/scoring decisions. These examples also mean we can walk through the different measurement decision stages and, hopefully, begin to demystify them. As we show in our analyses, the decisions researchers make about how to calibrate and score the survey used has consequences that are often overlooked, with likely implications both for conclusions drawn from individual psychological studies and replications of studies.

Translational Abstract

Considerable effort is often put into designing psychological studies, with great attention paid to various aspects of research design. However, when surveys are used to measure the outcome of interest, the approach used to score the survey is usually an afterthought. Measurement may be given short shrift because researchers wrongly assume that random noise due to measurement error is the main worry, or that nonrandom error will simply wash out between control and treatment groups. Alternatively, ignoring measurement may occur not because researchers make those assumptions, but because scoring models are complicated and related decisions labyrinthine. Whatever the reason, in most studies, people simply add up all the item responses to produce a sum score. When a sum score is not used, researchers usually assume that a garden variety item response theory (IRT) model is the main alternative. In this study, we show that, not only are there likely better scoring options available for common study designs like randomized control trials and growth/developmental studies; but using scoring approaches that do not match the study design, even when IRT is used, can severely bias results. To help researchers avoid such bias, we try to demystify the scoring process by walking through its decision stages, showing how decisions can affect study results, and providing code so that researchers can score their surveys in defensible ways.


Keywords: measurement, randomized control trials, treatment effects, interventions, item response theory (IRT)


Supplemental materials: <https://doi.org/10.1037/met0000506.supp>

Considerable effort is often put into designing psychological studies, with great attention paid to various aspects of research

design. For example, careful thought is often put into choice of experimental or quasi-experimental methods, sample recruitment, and analytic decisions. Researchers also often conduct power analyses to ensure that they will be able to detect a true treatment effect. This effort typically goes into large-scale studies with big budgets and small-scale studies alike.

However, the measurement model and scoring approach employed are often (though not always) an afterthought (Flake & Fried, 2020). For example, Flake et al. (2017) reviewed a representative sample of articles (using 433 survey scales) published in the *Journal of Personality and Social Psychology* for validity evidence supporting their

James Soland  <https://orcid.org/0000-0001-8895-2871>

Megan Kuhfeld  <https://orcid.org/0000-0002-2231-5228>

Correspondence concerning this article should be addressed to James Soland, School of Education and Human Development (EHD), University of Virginia, 405 Emmet Street, Charlottesville, VA 22904, United States. Email: Jgs8e@virginia.edu

intended uses. Roughly half of the scales included no citation to a prior validation study, “appearing to have been developed on the fly” (Flake et al., 2017, p. 374). Further, for half the scales, internal reliability was the only psychometric evidence provided and 19% of scales had no psychometric information whatsoever. This finding was corroborated by evidence that roughly 30% of education studies in the What Works Clearinghouse on literacy and science, technology, engineering, and mathematics (STEM) relied on outcomes derived from researcher-developed measures (Wolf, 2021). These studies provide some evidence that measurement is often not taken as seriously as other design and implementation issues.

One possible reason that measurement gets downplayed is that the story may seem clear: Measurement error in the independent variable will attenuate regression coefficients, and measurement error in the dependent variable will increase standard errors, but not otherwise introduce bias (Williams et al., 1995; measurement error in both the predictor and criterion can attenuate correlation coefficients as well). Additionally, prior research has shown that sum scores and scores from a more complicated psychometric model are often highly correlated with each other (Fan, 1998). However, this simple story obfuscates many other ways that measurement decisions might impact subsequent analyses. For example, the choice of measurement model, uncertainty in those parameters, and the calibration/scoring approaches could all impact whether estimated differences in the outcome of interest are biased or, even if unbiased, found to be significant (e.g., Cai, Choi, & Kuhfeld, 2016; Kuhfeld & Soland, 2020; Soland 2021).

Another possible reason is that using a two-step approach to scoring whereby a latent variable model (typically rooted in item response theory or IRT) is used to score the measure, then those scores are used in subsequent analyses, can be complicated. (This approach differs from, yet is directly related to, fitting measurement and structural models in tandem, as is done in structural equation modeling [SEM].) Even for researchers who worry about the consequences of measurement decisions, producing scores is a multistage and, oftentimes, confusing process. For example, one must decide whether to use a measurement model or not (e.g., using a sum score vs. IRT), what type of measurement model to fit, how best to calibrate the item parameters, then how to produce the scores. Even for seasoned measurement experts, these stages of decision-making can feel labyrinthine, and so researchers often fall back on the simplest available approaches (e.g., sum/mean scores). Little research attempts to cohesively explain those steps when scoring surveys, especially their consequences for recovering the estimand of interest.

In this study, we provide three motivating examples where surveys are used to understand individuals’ underlying social emotional and/or personality constructs to demonstrate the potential consequences of these measurement decisions. Our study focuses on short survey measures because they are more likely to be scored using sum scores, and because they are commonly used in psychological and educational studies (Gehlbach & Hough, 2018; McNeish & Wolf, 2020). In the first example, a survey is administered to a single group of individuals at a single timepoint, and we are primarily concerned with recovering the true score distribution (e.g., reporting a mean and standard deviation [*SD*] for a psychological construct). In the second example, a survey is administered at a single timepoint to multiple groups of individuals, and we are primarily interested in recovering the true difference in scores among groups (e.g., male-female gaps in self-efficacy). The final example involves a multigroup multi-time-point context, a design

widely seen in the randomized control trial (RCT) literature, where a survey is administered to a treatment and control group prior to and following an intervention. In addition to allowing us to show how measurement decisions can impact results from common study designs, these examples also mean we can walk through the different measurement decision stages and, hopefully, begin to demystify them. An additional goal is to help researchers understand when using IRT-based approaches may have negative consequences, such as when sample sizes are too small or survey scales are too short to produce reasonable scores, an issue examined extensively in other research (e.g., Sahin & Anil, 2017).

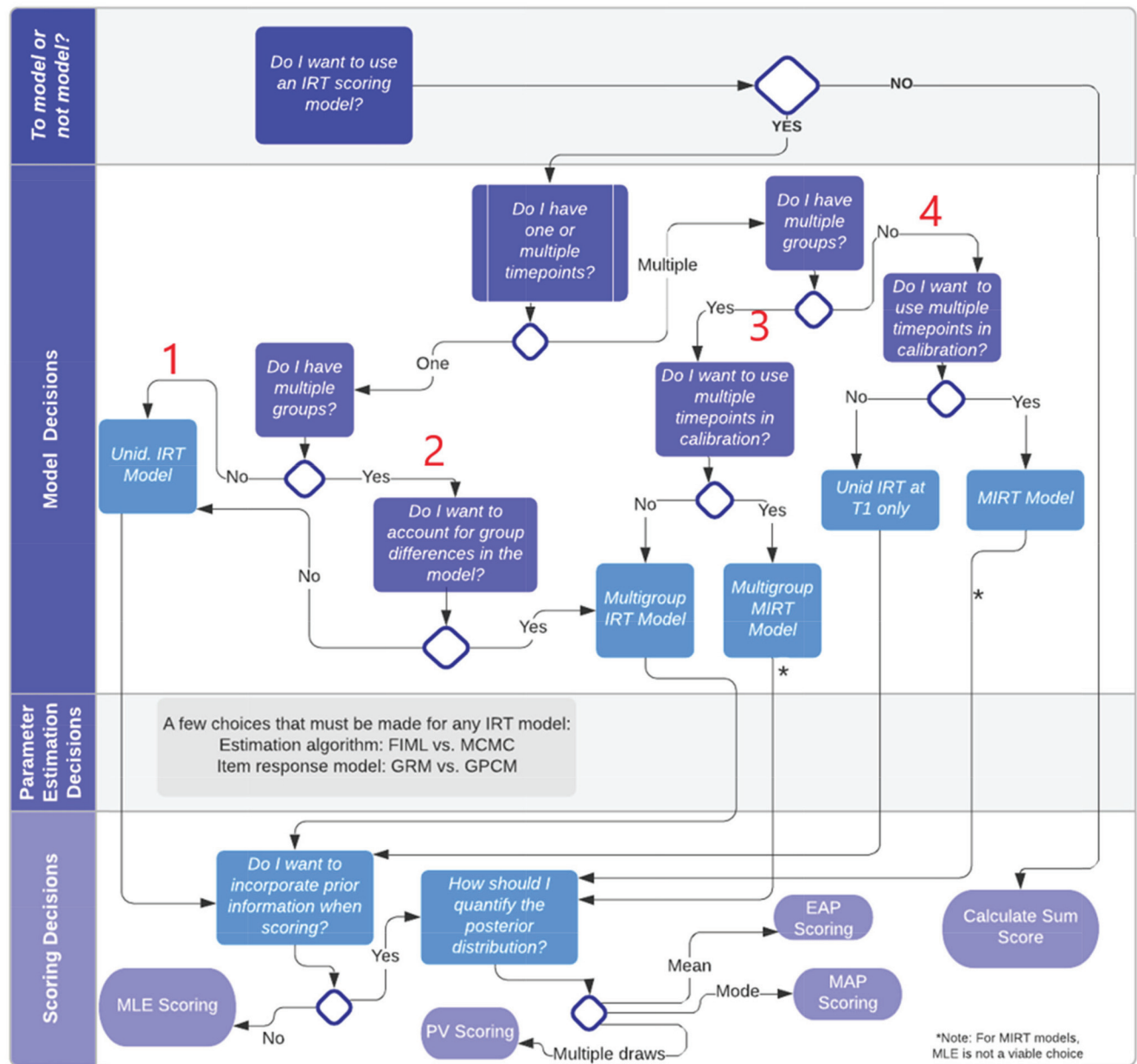
The study proceeds as follows. First, we describe one of the three examples, laying out the most common calibration and scoring approaches available. While we stay fairly conceptual in the main text, equations and code for the most relevant IRT models are provided in the online supplemental materials. Then, following each example, we conduct a simulation study examining the effects of the various decisions made. That is, rather than provide all background on scoring decisions at once, we instead describe the example then immediately follow with the associated simulation study before moving on to the next example. After walking through examples and corresponding simulations, we use empirical data to examine tradeoffs for the most complicated of the examples, namely a pre/post RCT design. Finally, we discuss implications and help researchers make informed decisions about scoring in their own studies. As we show in our analyses, the decisions researchers make about how to calibrate and score the survey used has consequences for each scenario (for readers who want a preview of the problems that can introduce bias and tradeoffs involved in making various decisions, they can skip ahead to Table 7). Further, these consequences are often overlooked, which likely has implications both for conclusions drawn from individual psychological studies and replications of studies (i.e., impacting the replication “crisis” in psychology).

Decision Stages Involved in Scoring

Before describing the motivating examples, we want to provide an overview of the stages of decision-making that a researcher must go through when producing scores based on a survey (throughout this study, we presume the work of designing and validating the survey for an intended purpose has been completed and the survey data have been collected). Figure 1 presents a flowchart of decisions, separated by four major stages. These stages are demarcated by the labels on the left-hand side of the figure and colored horizontal bands corresponding to those labels. The first stage (“To Model or Not Model” in the figure) requires the choice of whether to produce sum scores or use an IRT model. If sum scores are chosen, the decision process ends there. If the researcher chooses to produce scores from an IRT model,¹ then the second stage (“Model Decisions”) involves measurement modeling decisions, namely which calibration model to use (e.g., unidimensional, multidimensional, multigroup). The third stage (“Parameter Estimation Decisions”) involves decisions like which specific IRT

¹ We focus on IRT scoring models over other approaches for producing factor scores from factor models (such as Bartlett scoring with Croon’s correction) as IRT is more widely used in educational and psychological research to produce scores from latent variable models. For a review of factor score regression, see Devlieger et al. (2016).

Figure 1
Flowchart of Scoring Decisions



Note. The four stages of decisions are noted on the left side of the figure, while the various paths depicted in our motivating are marked in the figure with large red numbers. IRT = item response theory; MIRT = multidimensional item response theory; MLE = maximum likelihood estimation; PV = plausible values; EAP = expected a posteriori; MAP = modal a posteriori. Note that while multiple model decisions end at a single scoring location, this does not imply that the scores obtained from say a unidimensional IRT model will be equivalent to those from a MIRT model. See the online article for the color version of this figure.

approach should be used to implement the measurement model, which estimation algorithm to use,² and whether the sample size is sufficient to recover item parameters. Finally, the fourth stage (“Scoring Decisions”) involves selecting a scoring approach. As discussed in more detail below, our motivating examples reflect various paths through the flowchart.

² While we discuss in depth the measurement modeling and scoring approaches, we will not focus heavily on the various parameter estimation decisions. We refer readers to Wirth and Edwards (2007), Edwards (2010), to learn more about the tradeoffs of using full-information maximum likelihood or Markov chain Monte Carlo (MCMC) and to van der Linden (2018) for a review of the various item response models that can be fit to polytomous item response data.

Finally, note that, as we move through specific studies below, our simulations do not always explore all of the possible decisions in Figure 1. In some cases, that choice was made because other simulations have examined similar issues, such as how scoring decisions affect growth estimates in a single-group multiple timepoint situation (e.g., Bauer & Curran, 2016; Kuhfeld & Soland, 2020). In other cases, we do so to avoid overwhelming readers. For example, plausible values scoring is used in large-scale achievement tests (such as the National Assessment of Educational Progress [NAEP]) and is a viable option for scoring psychological measures, but is rarely used with surveys. Therefore, we do not discuss it in detail, nor include it in simulations.

Study 1: Scoring Individuals at One Timepoint

Decision Stages for Study 1

Our first study involves a short survey that has been administered at a single timepoint to a group of individuals with the intention of accurately capturing the true variability in the underlying latent trait (say, e.g., self-efficacy). Below, we discuss decision stages (corresponding to those in the flowchart in Figure 1) related to whether one should use a measurement model, which measurement models are plausible if the researcher decides to use one, options available for (and considerations related to) parameter estimation, and how to score the survey based on the measurement model. While the decisions listed below are not exhaustive, they likely represent the most plausible ones that researchers might make.

Stage I. Whether to Use a Measurement Model

Oftentimes, when using survey measures in psychology, researchers opt not to employ a measurement model, but instead to simply sum or average the item responses to produce a score. This approach has the advantages of being simple to implement, easy to replicate across studies, and has no minimum sample sizes or advanced software required. However, the simplicity of the choice to use sum scores masks the fact that this approach does in fact make strong assumption about the relationship between the items and the latent dimension of interest. For example, McNeish and Wolf (2020) showed that a sum score is equivalent to fitting an extremely constrained confirmatory factor analytic model. Such a model assumes that all the loadings across items are equal (and oftentimes set to one), and that the residual variances are also equal across items. Perhaps unsurprisingly, sum score use can lead to severe bias in observed scores, which can in turn result in misclassification of respondents (e.g., psychological patients receiving diagnoses; McNeish & Wolf, 2020). The impact of using a sum score could also differ dependent on the length of the survey scale (with longer surveys generally being more reliable, thus making sum scores potentially less problematic from a reliability standpoint) and how closely a less constrained version of the measurement model resembles the highly constrained sum score version.

Stage II. Modeling Decisions

If a researcher decides not to use sum scores, then a measurement model must be chosen. For the single group, single timepoint example, there is really one main option. However, these decisions become much more complex in subsequent examples.

If a researcher is only using a single timepoint and a single group, there is generally only one commonly used IRT model, namely the

unidimensional IRT model. In contrast to the sum score approach, there are several ways that item parameters could be estimated and scored in an IRT context (Gorter et al., 2016; Thissen & Wainer, 2001). A standard unidimensional model is shown in Figure 2, Panel A. The boxes (labeled y_1, \dots, y_n) represent the n observed item responses while the circle represents the unobserved latent characteristic θ . The strength of the relationship between the observed item responses and θ is represented by the item slope parameters (a_1, \dots, a_n). Assuming item j has K response options, there are $K-1$ item intercept parameters ($c_{j1}, \dots, c_{j,K-1}$), which are associated with the difficulty of the item. As is standard practice in unidimensional IRT modeling, we assume that θ is normally distributed with a mean of 0 and standard deviation of 1.

Stage III. Parameter Estimation Decisions

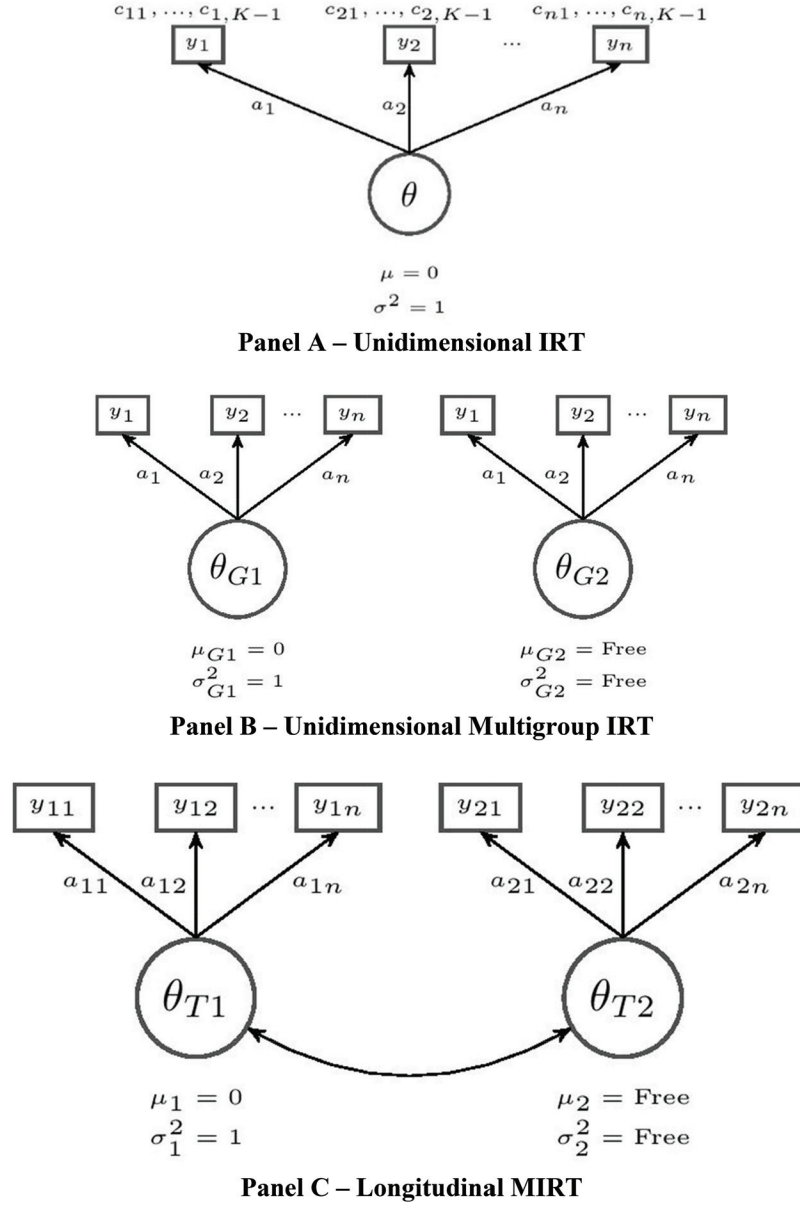
As noted above, because thorough comparisons of the impact of decisions related to parameter estimation exist elsewhere (e.g., Edwards, 2010; Wirth & Edwards, 2007), we focus primarily on highlighting the tradeoffs of the model and scoring decisions in this study. As result, we have chosen to hold the calibration decisions (graded response model [Samejima, 1969] using marginal maximum likelihood estimation via [Bock & Aitkin's, 1981] expectation-maximization [EM] algorithm) constant across our simulation conditions. For a thorough review of IRT models for ordinal response data and parameter estimation approaches, see Volumes 1 and 2 of van der Linden (2018).

No matter which estimator is used, uncertainty in the item parameters can be a concern in producing scores from an IRT model, especially when the study's sample size is small (Yang et al., 2012). A sum score essentially avoids estimating item parameters and therefore has no uncertainty in the parameter estimates, even if they are wrong. By contrast, IRT models estimate slope and intercept item parameters (among others). Issues of minimum sample size have been discussed elsewhere (e.g., Sahin & Anil, 2017), and should be kept in mind. Sample size, test length, and measurement model all matter in terms of item parameter recovery; Sahin and Anil (2017) showed that required sample sizes to recover parameters for dichotomous items ranged from 150 to 750 dependent on test length and model used.

Stage IV. Scoring Decisions

Maximum Likelihood Estimation Scoring Approach. After estimating item parameters in the IRT calibration step, analysts then need to decide how to estimate individual scores from the item response data. One approach sometimes used in standardized educational assessments is maximum likelihood estimation (MLE). Desirable properties of MLE are that it is asymptotically unbiased and that its standard error is related to the information function (Baker, 1992). Drawbacks of MLE, however, include infinite estimates for survey respondents whose response patterns use only the top or bottom category of the Likert scale (in plain terms, those individuals do not receive a score without additional action). To address this limitation, highest possible scale scores (HOSS) and lowest possible scale scores (LOSS) are often predefined before scoring for individuals with infinite scores, though the choice of which HOSS/LOSS values to use can be a bit arbitrary and potentially impact score variability. For example, if a survey respondent only uses the top response categories across all the items, his or her score would be undefined. To provide that person with a score, one could simply assign such individuals a

Figure 2
Path Diagram of Calibration/Scoring Approaches



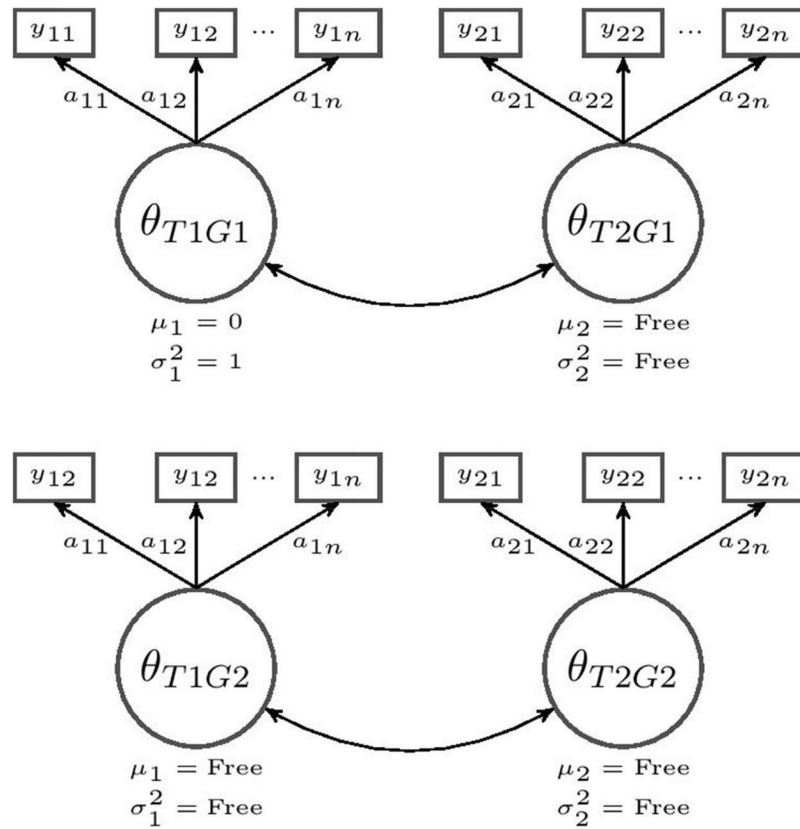
Note. IRT = item response theory; LM-MIRT = longitudinal multigroup - multidimensional item response theory. (Figure continues on next page)

score of, say, 3 SDs above the mean. Alternatively, one could simply treat such responses as missing. We examine both options and their implications in the simulation corresponding to Example 1.

Bayesian Scoring Approach. Meanwhile, Bayesian methods such as expected a posteriori (EAP) and maximum a posteriori (MAP) scoring do not share such limitations. Bayesian methods incorporate information about the population through the specification of prior distribution to approximate the posterior distribution of latent proficiency (Bock & Mislevy, 1982). Under the Bayesian paradigm, the posterior distribution of the proficiency levels (i.e., θ) is defined as

the product of the likelihood function and the prior distribution. The mean of the posterior distribution is the proficiency estimate under EAP, whereas the mode of the posterior distribution is the proficiency estimate under MAP (Yen & Fitzpatrick, 2006). The choice of a reasonable prior distribution for the proficiency level is key to Bayesian estimators. The most common prior distribution is the standard normal distribution, $N(0, 1)$, though it is also possible to incorporate covariates information into the prior distribution (we discuss this further in the following example). EAP and MAP scoring approaches are shrinkage estimators that shrink the latent distribution toward the

Figure 2. (continued)



Panel D – LM-MIRT

population average, though the degree of shrinkage depends on the test length and reliability (Thissen & Orlando, 2001). That is, the approach accounts for some uncertainty in the scores by shrinking them to the mean/mode based on the level of uncertainty.

Plausible Values Scoring Approach. Though not examined in this study, an additional scoring approach that is widely used in large-scale assessments such as the National Assessment of Educational Progress (NAEP) is to draw plausible values (PV) from the posterior distribution. PV are draws from the posterior distribution, which is similar to multiple imputation in a missing data context, and better conveys the uncertainty associated with the construct of interest than a single point estimate (von Davier et al., 2009). However, such an approach is, so far as we can tell based on a brief review of the literature, used much less when scoring surveys.

Simulation Study 1

Next, we conduct a simulation study to help quantify the effects of related decisions when the study design involves a single group at a single timepoint. Again, readers who want to preview the

results for the simulation studies can turn to Table 7. Table 1 includes details on simulation conditions across examples/studies.

Data Generation and Scoring

The first simulation study examined the sensitivity of scoring approaches in a simple scenario where all simulees were drawn from a single population. As detailed in Table 1, we varied the survey length (four, eight, and 12 items) and sample size (100–1,000 individuals). We also varied the relationship between the items and the latent dimension (item slope parameters) such that the scale itself was more or less reliable (beyond differences in reliability dictated by the length of the scale, with longer scales generally being more reliable). All generating item parameters can be found in Table 2. As the table shows, high reliability³ slope parameters were produced based on survey scales shown using empirical data to have high loadings in a confirmatory factor analytic framework (Soland, 2021), and moderate

³ The “moderate reliability” condition had alpha values of .7 for the four-item measure, .85 for the eight-item measure, and .9 for the 12-item measure. The “high reliability” condition had alpha values of .8 for the four-item measure, .9 for the eight-item measure, and .95 for the 12-item measure.

Table 1
Simulation Conditions by Study

Conditions	Study 1	Study 2	Study 3
Groups	1	2	2
Timepoints	1	1	2
Measurement model	IRT	IRT, multigroup MIRT	IRT, MIRT, LM-MIRT
Scoring approach	EAP, MLE	Sum, EAP, MLE	Sum, EAP
MLE treatment of high/low scores	Missing, replaced with max/min	Missing, replaced with max/min	N/A
Survey length	4, 8, 12 items	4, 8, 12 items	4, 8, 12 items
Item reliability	High, Moderate	Moderate	Moderate
Sample size	100, 200, 300, 400, 500, 1,000	100, 150, 200, 250, 500, 1,000 per group	100, 150, 200, 250, 500, 1,000 per group
True mean/mean difference	Mean = 0	Mean difference = 0, .1, .25 SDs	Mean difference at T2 = 0, .1, .25 SDs

Note. IRT = item response theory; MIRT = multidimensional item response theory; MLE = maximum likelihood estimation; EAP = expected a posteriori.

reliability slope parameters were produced by reducing those parameters uniformly by .5 units.

We compared sum scores with multiple IRT-based scoring approaches based on a unidimensional graded response model.⁴ Specifically, we produced scores with a unidimensional IRT model using MLE and EAP. Further, for MLE, when a simulated individual used only the top or bottom response category, we both treated that score as missing ("MLE missing") and replaced it with a maximum/minimum score (we used the software default of $-4/4$ SDs, and refer to this condition as "MLE maximum"). We also compared to true scores produced during the data generation process. These true scores allow us to see, for example, how often Type I errors might occur when using a survey respondent's true position on the latent continuum, and compare that rate to the one produced by IRT models designed to estimate that true score. All conditions were replicated 500 times in flexMIRT v3.51 (Cai, 2017).

Finally, once scores were produced, we compared their means and variances by condition to see how scoring decisions affected those parameter estimates. We also examined bias, defined as the difference between the parameter estimate (averaged across replications) and the true (data-generating) parameter.⁵ Sum scores were not used because scaling differences would affect the means and variances for the scores such that they are on different scales. Further, placing them on the same scale as one of the IRT-based scores would mainly serve to

make the bias in the means and variances match that of the IRT model in question. However, we use sum scores in subsequent simulation studies when comparing Type II error rates.

Results

Table 3 presents the means and variances (SDs) across scoring approaches by sample size, survey length, and marginal reliability of the measure. Beginning with the means, MLE maximum scores were often quite biased, and upwardly so (note that bias is relative to the data-generating parameter, not the estimate using true scores, which can themselves produce slightly biased estimated means). This result occurred because, as is often the case with survey scales, the top categories were used much more frequently than the bottom, which meant there were more cases where scores were set at the maximum of 4 SDs. This bias diminished as the length of the survey increased given the probability of using only a single response category decreased as respondents saw more items. The effect of this bias was reduced when the items were more reliable. Meanwhile, the MLE missing scores were also biased, though not nearly as much as with MLE maximum. The bias was in the opposite direction because, whereas replacing extreme responses with the maximum inflated scores, treating those same scores as missing deflated the mean. By contrast, EAP scores showed no bias out to three decimal places. In fact, EAP scores showed less bias in recovering the true mean parameter ($\mu = 0$) than when the mean was calculated using the generating (true) scores.

This somewhat counterintuitive result between EAP and true scores relates to score shrinkage. The SD of the EAP scores was well under the true SD of 1, while the SD of the MLE scores was well over 1. EAP scores tended to understate the variance due to shrinkage, with scores from shorter, less reliable survey scales being shrunk much closer to the prior distribution mean. Notably, while the length of the survey scale heavily affected shrinkage, reliability of the items had relatively minimal effect, and sample size even less effect. By contrast, MLE scores included no shrinkage, producing variances that were biased upward. To help make this contrast clearer, Figure 3 plots SDs of EAP scores against those of MLE maximum scores, with different

Table 2
Generating Item Parameters for Both Groups, Four-Item Scale

Item	a1	c1	c2	c3	c4
1	0.75	2.13	0.37	-0.86	-1.88
2	1.30	2.67	1.33	0.31	-0.67
3	1.80	3.57	1.70	0.07	-1.90
4	1.65	3.77	2.32	1.09	-0.31
5	0.94	3.38	0.68	-0.77	-1.60
6	1.85	3.17	1.50	0.37	-0.74
7	2.53	3.71	3.00	0.06	-3.11
8	2.28	3.95	2.86	1.19	-0.54
9	1.18	3.26	0.59	-0.97	-1.92
10	1.76	2.74	1.14	0.33	-0.54
11	1.59	3.66	1.55	0.07	-2.46
12	1.45	3.74	2.73	0.88	-0.36

Note. For conditions with only four or eight items, only the first four/eight items in the table are used. For studies with multiple timepoints and/or groups, measurement invariances is assumed across timepoints and groups. When there are high/low loading conditions, the a1 column was changed uniformly as described in the text. Note that we report intercepts rather than thresholds in the table.

⁴ We used an approach detailed by Kolen and Brennan (2014) to attempt to place sum scores on the same scale as the IRT scores to improve comparability. However, doing so only made the means/variances match those of the selected IRT scores, and there was no effect on Type I/II errors. Therefore, we do not report these results.

⁵ Item parameter recovery and bias is presented in Table 3.1 in the online supplemental materials. Significant bias was observed for $N = 100$, but not the other sample sizes.

Table 3*Simulations Results for Single Group, Single Timepoint Simulations*

Moderate reliability		Means				SDs			
<i>N</i>	Items	EAP	MLE (max)	MLE (miss.)	True score	EAP	MLE (max)	MLE (miss.)	True score
100	4	0.000	0.101	−0.024	−0.014	0.850	1.399	1.172	1.003
200	4	0.000	0.089	−0.032	−0.016	0.836	1.388	1.175	0.999
300	4	0.000	0.089	−0.030	−0.011	0.835	1.382	1.173	1.004
400	4	0.000	0.086	−0.029	−0.009	0.831	1.381	1.173	1.004
500	4	0.000	0.088	−0.032	−0.006	0.832	1.381	1.174	1.005
1,000	4	0.000	0.086	−0.037	−0.006	0.829	1.381	1.170	1.005
100	8	0.000	0.040	0.004	0.009	0.930	1.184	1.102	0.993
200	8	0.000	0.040	0.002	0.009	0.926	1.181	1.102	0.994
300	8	0.000	0.039	0.001	0.004	0.926	1.179	1.101	1.001
400	8	0.000	0.037	0.000	0.003	0.924	1.174	1.094	1.000
500	8	0.000	0.036	0.001	0.001	0.924	1.170	1.100	0.998
1,000	8	0.000	0.037	0.002	0.001	0.924	1.172	1.097	0.999
100	12	0.000	0.022	0.007	0.000	0.951	1.113	1.076	1.007
200	12	0.000	0.021	0.008	−0.011	0.947	1.106	1.077	1.004
300	12	0.000	0.022	0.007	−0.008	0.946	1.110	1.077	1.001
400	12	0.000	0.022	0.008	−0.008	0.945	1.108	1.079	0.999
500	12	0.000	0.021	0.008	−0.005	0.945	1.108	1.078	1.001
1,000	12	0.000	0.022	0.008	−0.004	0.944	1.108	1.078	0.997

High reliability		Means				SDs			
<i>N</i>	Items	EAP	MLE	MLE (miss.)	True score	EAP	MLE	MLE (miss.)	True score
100	4	0.000	0.073	−0.035	−0.014	0.897	1.291	1.078	1.003
200	4	0.000	0.065	−0.030	−0.016	0.890	1.273	1.086	0.999
300	4	0.000	0.066	−0.032	−0.011	0.889	1.271	1.082	1.004
400	4	0.000	0.065	−0.031	−0.009	0.887	1.274	1.080	1.004
500	4	0.000	0.067	−0.033	−0.006	0.886	1.273	1.084	1.005
1,000	4	0.000	0.065	−0.033	−0.006	0.886	1.272	1.077	1.005
100	8	0.002	0.035	−0.008	0.009	0.952	1.144	1.053	0.993
200	8	0.000	0.031	−0.008	0.009	0.949	1.134	1.055	0.994
300	8	0.000	0.031	−0.011	0.004	0.949	1.136	1.051	1.001
400	8	0.000	0.031	−0.012	0.003	0.948	1.135	1.049	1.000
500	8	0.000	0.030	−0.010	0.001	0.948	1.129	1.050	0.998
1,000	8	0.000	0.031	−0.012	0.001	0.947	1.132	1.049	0.999
100	12	0.001	0.019	0.000	0.000	0.968	1.085	1.042	1.008
200	12	0.000	0.018	−0.001	−0.011	0.965	1.081	1.042	1.004
300	12	0.000	0.018	−0.002	−0.008	0.964	1.082	1.040	1.001
400	12	0.000	0.018	−0.002	−0.008	0.963	1.082	1.038	0.999
500	12	0.000	0.018	−0.002	−0.005	0.963	1.082	1.040	1.001
1,000	12	0.000	0.018	−0.002	−0.004	0.962	1.082	1.037	0.997

Note. MLE = maximum likelihood estimation; EAP = expected a posteriori.

marker symbols for item slopes and survey scale length (there are clusters of points rather than a single point for each combination because of the five different sample sizes used). As item slopes and scale length decreased, EAP variances decreased and MLE variances increased. In general, changes in variances were more pronounced by survey length than by item slopes.

Though not reported, correlations among scores produced using these various approaches are highly correlated, including with true scores from the data-generating models (the survey respondent's actual position on the latent continuum that all of these scoring approaches are essentially trying to recover). Specifically, for the 12-item survey, all scores (IRT-based and sum) are correlated with true scores at approximately .94. Thus, purely examining how students are ordered based on the different scores—a common practice when evaluating the effects of scoring—would mask other issues of consequence to recovering true means and variances.

All told, using MLE tends to produce biased estimates of the means (especially for shorter survey scales using MLE maximum)

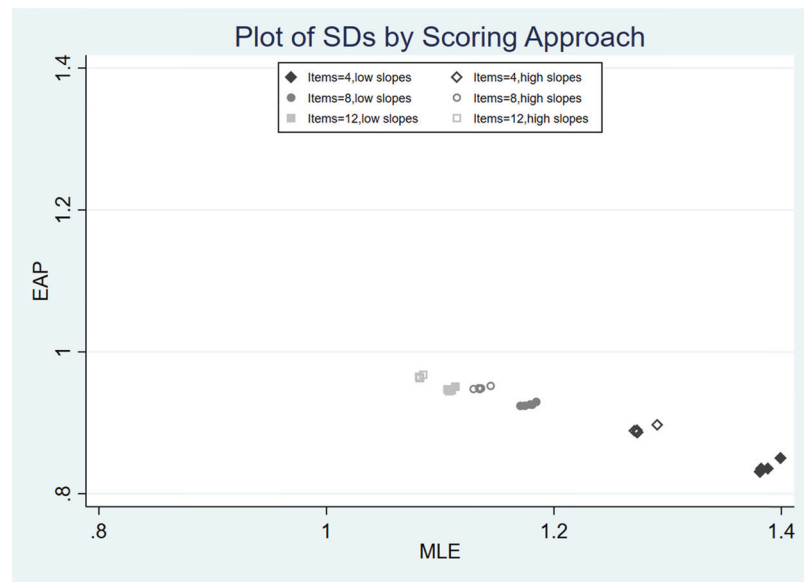
and, because it does not shrink scores to address unreliability, produced variances that were biased upward. By contrast, EAP scores produced unbiased means, in part because scores are shrunken considerably when using a short survey scale. However, variances were downwardly biased. These results—especially differences in variances by scoring approach—would surely effect group comparisons, thus setting the groundwork for the next example and simulation study comparing group means.

Study 2. Calibration and Scoring With Multiple Groups at One Timepoint

Decision Stages for Study 2

In the second example we examine, two or more groups filled out a survey at a single timepoint, and we are interested in mean differences between their scores. For instance, a researcher may be interested in

Figure 3
Plot of Score SDs, Single Timepoint Single Group Simulation



Note. Within each survey length and slope condition, the average *SD* is displayed for each of the five sample size conditions. See the online article for the color version of this figure.

gender differences in student-reported self-efficacy in college students. Similarly, a researcher might have conducted an RCT comparing means from two groups following an intervention.

Stage I. Whether to Use a Measurement Model

As in all our examples, one could simply sum or average the item responses to produce a score. And, as before, doing so makes very strong assumptions, including that the association between the latent variable and the observed item responses is the same across items, and that the residual variances are equal. These assumptions can lead to meaningful bias, including when comparing means (Soland, 2021). However, in the multigroup scenario, there are additional considerations. For example, using sum scores complicates if not obviates testing for measurement invariance across groups. Further, one cannot employ a partial measurement invariance model by which certain invariance assumptions are relaxed, such as allowing the intercept parameter for a given item to differ across groups. Ignoring measurement noninvariance between groups, including control and treatment groups, can severely bias estimates of interest (e.g., treatment effect estimates; Soland, 2021).

Stage II. Modeling Decisions

Unidimensional IRT Model. If a researcher wants, a simple unidimensional IRT model can once again be used. Like the sum score, such a model does not generally allow for a partial invariance model. However, one could still test for noninvariance across groups in an IRT context before concluding it is safe to proceed with consistent parameters across groups. As another limitation of this model, using the unidimensional model in this case assumes that there is a single population

(typically one that follows a standard normal distribution), whereas multigroup IRT models described next assume that each group has its own latent mean and variance. We discuss the implications of this assumption more below.

Multigroup IRT. One could further expand the unidimensional IRT model to fit a multigroup IRT model, which allows for additional examination of measurement invariance across groups and equality of the latent mean and variance across groups (see Figure 2, Panel B). Such models parallel the single-group IRT model but allow for both (a) measurement noninvariance across the groups in the calibration model and (b) different mean and variance structures between the two groups. For example, if we are comparing boys' and girls' self-efficacy, "the boys'" latent distribution can be scaled to a standard normal distribution and "the girls'" latent mean and variance can be freely estimated in the multigroup IRT model. Thus, even before scores are produced, population-level group differences can be estimated in the IRT model. This approach is akin to the one used in SEM where no scores are produced; rather, the measurement model and differences in latent means/variances across groups are estimated all in one step.

Unidimensional IRT With Latent Regression. Though not the focus of our own study (and thus not part of our simulations), one might be interested in comparing more than one group, or even accounting for differences in survey respondents that are continuous in nature when calibrating and scoring. In such cases, one could also calibrate with a unidimensional IRT model that incorporates information on background characteristics (i.e., covariates) when generating scores. For example, Curran et al (2018) found that using covariate-informed factor score estimates substantially mitigates bias in subsequent analyses involving the factor scores

and covariates. In the NAEP context, conditioning information is also incorporated into the measurement and scoring model so that estimates from the student assessments are statistically consistent for the population comparisons of interest (Mislevy et al., 1992).

Stage III. Parameter Estimation Decisions

We do not consider these decisions in depth here because they largely parallel the same decisions in Example 1. As discussed more in the third example, when model complexity increases (e.g., going from the single group to the multigroup IRT model), the sample size needed to fit an IRT model also generally increases.

Stage IV. Scoring Decisions

As in the first example, one could use MLE scoring. The same tradeoffs are generally involved here. However, the implications of using EAP/MAP are different in the multigroup context. In most examples of EAP/MAP scoring using a unidimensional model, only students' item responses and a single population distribution are used to produce student scores. Thus, there is only one aggregate population mean and variance. As a result, scores from both groups are shrunk to a single mean. What does this choice mean practically? In the context of most experiments where the outcome of interest is measured via a test or survey and a single-group IRT model is used for scoring, scores are produced assuming full exchangeability of subjects across treatment and control groups (Lindley & Smith, 1972).

By contrast, when a multigroup IRT model is used with EAP/MAP scoring, population distributions appropriate to each group are used as the prior distribution, which avoids potential biases induced by shrinking both groups to a common mean. That is, the multigroup IRT model with EAP/MAP scoring does not assume exchangeable subjects. As our simulations will show, shrinking two groups being compared using one mean versus two has considerable implications for being able to detect true treatment effects.

Simulation Study 2

Data Generation

Simulation Study 2 is identical to Study 1 (including using the same item parameters as in Table 2 and assuming measurement invariance between groups), except now a multigroup IRT model is used to simulate item responses from two independent groups rather than assuming all individuals are drawn from a single population. For this data generation approach, true ability estimates were generated for group g (θ_g , where $g = 0, \dots, G - 1$), and we assume students are sorted into either a treatment or control group. Thus, there were two latent means of interest: $\mu_{g=0}$ (control group) and $\mu_{g=1}$ (treatment group), where μ_g is the true mean of θ_g . Both groups' scores had a variance of one, and $\mu_{g=0}$ was set to zero. In this study, the sample size (N) refers to the number of simulees per group, so the total sample sizes are twice as large as in Study 1.

Meanwhile, we varied $\mu_{g=1}$ to have means of 0, .1 *SD*, and .25 *SDs*. We chose these differences, which can be interpreted as effect sizes in units of the true scores, for a few reasons. First, we used a true group difference of zero such that significant mean differences would represent a spurious finding, and Type I error rates could be examined. Second, we chose .1 *SD* because effect sizes of that magnitude have been reported in other psychological

studies (Yeager & Walton, 2011), and because they are small enough that Type II errors could occur. Third, we used an effect size of .25 because it was both plausible and fairly large (Szucs & Ioannidis, 2017), yet still small enough that Type II errors still occur (especially when the sample size is small or model misspecified). While other, larger effect sizes occur with regularity in psychology (Szucs & Ioannidis, 2017), including them in the simulations neither changed the story about parameter recovery, nor provided any illumination on Type II errors (given most findings tended to be significant with an effect size of .25). Thus, we used the three we selected to tell the cleanest story while acknowledging that results could differ for other effect sizes.

Scoring

As shown in Table 1, all survey items were scored using sum, EAP, and MLE approaches.⁶ We also produced IRT-based scores using both single group and multigroup unidimensional models (and scored both using EAP and MLE). Finally, for these various scores, we examined mean differences, the proportion of significant mean differences (1 – Type II error) across replications, and the variances of the scores/differences. Specifically, when examining mean differences, we used two-group t tests to determine if there were significant differences between groups.

Results

Table 4 shows estimated mean differences between groups by scoring approach and simulation condition.⁷ When the true mean difference was zero, one can interpret the estimate as the average bias in the estimated mean difference. For example, when using shorter survey scales with small sample sizes, several of the scoring approaches tended to produce mean differences with a slight upward bias (above zero). Meanwhile, when the true mean difference was not equal to zero, one could obtain the bias by subtracting the true mean difference from the estimated. As the table shows, while all scoring approaches tended to produce some bias, it was most pronounced when using MLE maximum (upward bias), as well as both single group measurement models. Once again, bias tended to be worse at smaller sample sizes and for shorter, less reliable scales.

To help make these tradeoffs clearer, Figure 4 shows kernel density plots of the distribution of estimated mean differences across replications by scoring approach for the 12-item condition with a true treatment effect of .25 *SDs* (unlike in Table 4, sum scores are included here as a point of comparison despite scaling differences). As the figure shows, considerable bias tends to be introduced into estimated mean differences (e.g., differences in means between control and

⁶ Sum scores were once again rescaled to match the scale of the EAP and MLE scores using the approach detailed by Kolen and Brennan (2013). Though not discussed in the main text, another viable approach to attempt to make sum score results more comparable to those using other approaches is to produce a Cohen's d effect size, convert that effect size to a correlation, correct the correlation for attenuation, then convert back to an effect size. This approach was described by Schmidt and Hunter (2015). Results can be found in the supplemental online materials. In general, correcting for attenuation does tend to make effect sizes using sum scores more comparable to those based on true scores from the data generation, but certainly does not eliminate all discrepancies.

⁷ Item parameter recovery is displayed in Table 3.2 in the online supplemental materials. Parameters were accurately recovered in all but the $N = 100$ per group condition.

Table 4*Estimates of Mean Differences Between Groups, Two Group Single Timepoint Simulation*

Mean dif. = 0			Multigroup IRT model			Single group IRT model	
<i>N</i>	Items	True score	EAP	MLE (max)	MLE (miss.)	EAP	MLE
100	4	0.011	0.011	0.012	0.002	0.013	0.017
200	4	0.019	0.018	0.018	0.014	-0.017	-0.022
500	4	0.007	0.012	0.012	0.013	0.005	0.008
1,000	4	0.002	0.004	0.005	0.001	0.001	0
100	8	-0.010	-0.009	-0.01	-0.008	0.012	0.017
200	8	0.001	0.002	0.001	0.003	-0.012	-0.015
500	8	0.002	0.002	0.002	0.005	-0.007	-0.008
1,000	8	0.001	0.004	0.004	0.001	0.003	0.005
100	12	0.002	0.007	0.008	0.001	0.014	0.023
200	12	0.017	0.018	0.018	0.016	-0.008	-0.008
500	12	0.003	0.002	0.003	0.001	0.008	0.01
1,000	12	0.006	0.01	0.011	0.01	0.007	0.008

Mean dif. = .1			Multigroup IRT model			Single group IRT model	
<i>N</i>	Items	True score	EAP	MLE (max)	MLE (miss.)	EAP	MLE
100	4	0.111	0.111	0.123	0.096	0.057	0.089
200	4	0.119	0.118	0.128	0.104	0.04	0.049
500	4	0.107	0.113	0.125	0.103	0.065	0.078
1,000	4	0.102	0.103	0.116	0.089	0.047	0.074
100	8	0.090	0.093	0.098	0.094	0.057	0.089
200	8	0.101	0.102	0.108	0.102	0.044	0.056
500	8	0.102	0.102	0.109	0.102	0.052	0.061
1,000	8	0.101	0.102	0.11	0.098	0.059	0.075
100	12	0.102	0.111	0.117	0.103	0.06	0.099
200	12	0.117	0.118	0.123	0.117	0.048	0.063
500	12	0.103	0.102	0.106	0.101	0.067	0.079
1,000	12	0.106	0.111	0.115	0.11	0.066	0.078

Mean dif. = .25			Multigroup IRT model			Single group IRT model	
<i>N</i>	Items	True score	EAP	MLE (max)	MLE (miss.)	EAP	MLE
100	4	0.261	0.261	0.293	0.232	0.124	0.199
200	4	0.269	0.271	0.3	0.235	0.125	0.157
500	4	0.257	0.261	0.292	0.235	0.151	0.18
1,000	4	0.252	0.252	0.283	0.221	0.113	0.186
100	8	0.240	0.248	0.263	0.242	0.124	0.202
200	8	0.251	0.255	0.273	0.252	0.129	0.165
500	8	0.252	0.254	0.271	0.25	0.14	0.165
1,000	8	0.251	0.253	0.271	0.243	0.143	0.182
100	12	0.252	0.259	0.271	0.252	0.127	0.21
200	12	0.267	0.269	0.281	0.266	0.132	0.17
500	12	0.253	0.252	0.263	0.249	0.154	0.182
1,000	12	0.256	0.262	0.273	0.26	0.154	0.183

Note. IRT = item response theory; MLE = maximum likelihood estimation; EAP = expected a posteriori. *N* here refers to the number of simulees per group.

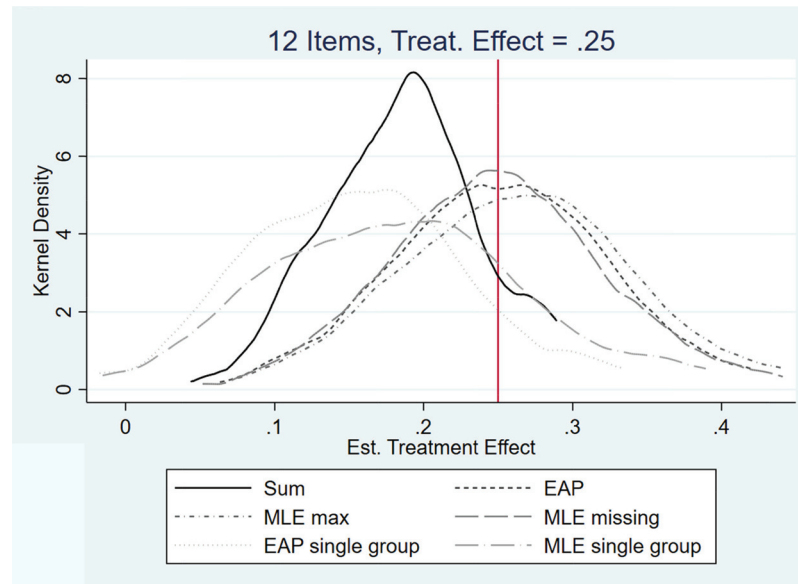
treatment groups), especially using the single group measurement model under either scoring approach. As we discuss momentarily, the downward bias in these estimates has nonnegligible impacts for Type II error rates. In short, using a measurement model that does not match the data-generating process can put researchers at a disadvantage when trying to detect true treatment effects.

Meanwhile, Table 5 shows the proportion of significant mean differences (using $\alpha = .05$) by scoring approach and condition. Thus, when the true mean difference is zero, one could interpret a given cell as the proportion of Type I errors by scoring approach across replications. As the table shows, when using shorter survey scales, Type I error rates are much higher for EAP than the other approaches. In some cases, EAP scores can produce Type I error rates that are more

than quadruple the rate from other scoring approaches. When the survey scale is longer (12 items), the Type I error rate for EAP scores was much smaller, though still above .05. Specifically, with 12 items, roughly 11% of the replications assuming a true mean difference of zero produced significant results using EAP.⁸ While this value is higher than one might like, even true scores tended to produce significant results when the true difference was zero at a rate above 5% (oftentimes in the 7% range), as do most of the other scoring

⁸To help ensure these higher rates were not simply due to sampling error, we ran these simulations using an additional 400 replications for a total of 500. While the proportion of significant results shifted slightly, general conclusions did not change.

Figure 4
Kernel Densities of Estimated Treatment Effects by Scoring Approach



Note. MLE = maximum likelihood estimation; EAP = expected a posteriori. EAP, MLE max, and MLE missing are all based on the multigroup IRT model. See the online article for the color version of this figure.

approaches including sum scores. In short, because group-specific *SDs* are smaller due to shrinkage for EAP scoring, Type I error rates for EAP scores tend to be higher than for other scoring approaches, and while having a longer (more reliable) survey scale mitigates this issue substantially, it is still a concern. Further, Type I error rates are a concern regardless of scoring approach.

Conversely, when the true mean difference is not 0, EAP scores tend to reduce Type II error rates considerably, including relative to sum scores. Due to shrinkage, EAP scores actually do a better job of reducing Type II errors than when using true scores. In general, non-EAP scoring approaches tend to produce fairly similar patterns of Type II error rates compared to true scores. To make such Type II (and Type I) error rates more concrete, Figure 5 shows the proportion of significant mean differences by scoring approach and measurement model (sample size of 100). As the figure shows, multigroup EAP scores almost always produce more significant results than multigroup MLE scores, including when the true mean difference is zero. However, when comparing to true scores, multigroup EAP scores tend to produce slightly more significant results when the true mean difference does not equal zero, but far more when the true mean difference does equal zero. That is, Type I error rates are much higher for multigroup EAP scores compared with both multigroup MLE and true scores. Note that this issue with multigroup EAP scores mainly arises when the survey scale consists of only four items. While Type I error rates remain higher for multigroup EAP scores when using a longer scale, the rates are not nearly as severe, and are much closer to Type I error rates for true and multigroup MLE scores.

Turning to single group models and sum scores, there are also implications for Type I and II errors. As the bottom row of the figure illustrates, single group EAP scores produce far fewer

significant results than multigroup EAP scores. In some cases, the proportion of significant results nearly quadruples. By contrast, multigroup EAP scores also produce far more significant results than when using sum scores. As before, while multigroup EAP scores also produce more Type I errors, the issue is problematic mainly when using only a very short (four item) survey scale.

Study 3. Calibration and Scoring With Multiple Groups at Multiple Timepoints

Decision Stages for Study 3

Finally, in our third example we consider a common RCT design with two groups (treatment and control) who are assessed prior to and following some intervention. Under this design, the researcher must choose whether to incorporate multiple timepoints and multiple groups into any calibration/scoring approach. Implications of such decisions are examined next.

Stage I. Whether to Use a Measurement Model

All of the assumptions made when using sum scores and the problems they can produce in the single timepoint single group example and the single timepoint multigroup example remain in the multigroup multi-timepoint example. However, even more assumptions are added in this context. Sum scores assume that the measure functions identically across timepoints, an assumption that is frequently violated in longitudinal studies that span developmental periods (Millsap, 2012). Additionally, the same items must be repeatedly administered for sum scores to be used to estimate change over time, which can be problematic if the period between assessments is short enough for individuals to remember their responses to the

Table 5*Proportion of Significant Treatment Effect Estimates, Two Group Single Timepoint*

N	Items	Multigroup IRT model			Single group IRT model		No model	
		EAP	MLE (max)	MLE (miss.)	EAP	MLE	Sum	True score
Mean dif. = 0								
100	4	0.190	0.030	0.060	0.102	0.092	0.040	0.060
200	4	0.270	0.070	0.060	0.041	0.041	0.070	0.100
500	4	0.160	0.020	0.030	0.082	0.071	0.020	0.030
1,000	4	0.170	0.020	0.070	0.072	0.062	0.030	0.040
100	8	0.090	0.070	0.060	0.041	0.041	0.050	0.060
200	8	0.040	0.010	0.020	0.031	0.041	0.020	0.030
500	8	0.150	0.080	0.070	0.143	0.133	0.060	0.030
1,000	8	0.140	0.080	0.070	0.052	0.021	0.090	0.100
100	12	0.080	0.040	0.040	0.071	0.041	0.050	0.060
200	12	0.110	0.060	0.050	0.051	0.041	0.070	0.070
500	12	0.110	0.070	0.050	0.082	0.102	0.050	0.050
1,000	12	0.110	0.050	0.070	0.062	0.062	0.070	0.070
Mean dif. = .1								
100	4	0.230	0.090	0.080	0.112	0.102	0.110	0.120
200	4	0.330	0.190	0.150	0.071	0.041	0.190	0.200
500	4	0.560	0.290	0.300	0.174	0.174	0.330	0.400
1,000	4	0.720	0.490	0.390	0.165	0.134	0.470	0.630
100	8	0.150	0.070	0.060	0.061	0.061	0.060	0.110
200	8	0.220	0.100	0.150	0.051	0.041	0.080	0.180
500	8	0.470	0.290	0.320	0.143	0.143	0.290	0.370
1,000	8	0.700	0.580	0.540	0.237	0.217	0.540	0.630
100	12	0.180	0.130	0.120	0.071	0.092	0.100	0.130
200	12	0.280	0.240	0.220	0.061	0.061	0.220	0.250
500	12	0.410	0.330	0.310	0.143	0.153	0.290	0.360
1,000	12	0.700	0.630	0.590	0.289	0.288	0.630	0.630
Mean dif. = .25								
100	4	0.570	0.320	0.250	0.174	0.184	0.310	0.490
200	4	0.790	0.540	0.470	0.316	0.337	0.540	0.750
500	4	0.990	0.930	0.850	0.602	0.551	0.930	0.980
1,000	4	1.000	1.000	0.990	0.784	0.784	1.000	1.000
100	8	0.420	0.290	0.330	0.112	0.122	0.300	0.390
200	8	0.730	0.600	0.640	0.245	0.225	0.620	0.750
500	8	0.970	0.950	0.910	0.520	0.520	0.930	0.980
1,000	8	1.000	1.000	1.000	0.825	0.825	1.000	1.000
100	12	0.500	0.390	0.410	0.184	0.184	0.420	0.430
200	12	0.760	0.700	0.670	0.204	0.194	0.670	0.730
500	12	0.980	0.950	0.950	0.604	0.602	0.960	0.970
1,000	12	1.000	1.000	1.000	0.907	0.905	1.000	1.000

Note. IRT = item response theory; MLE = maximum likelihood estimation; EAP = expected a posteriori. *N* here refers to the number of simulees per group.

items (particularly for assessments of content knowledge). However, IRT-based approaches allow for a wide range of equating and scaling methods to vary the format of the measure across timepoints while still allowing for comparable scores. These assumptions and limitations of sum scores have nontrivial implications for common uses of longitudinal scores, such as growth modeling. Kuhfeld and Soland (2020) showed that, under certain scenarios, using sum scores lead to an understatement of latent slope parameter means and variances by nearly 50%.

Stage II. Modeling Decisions

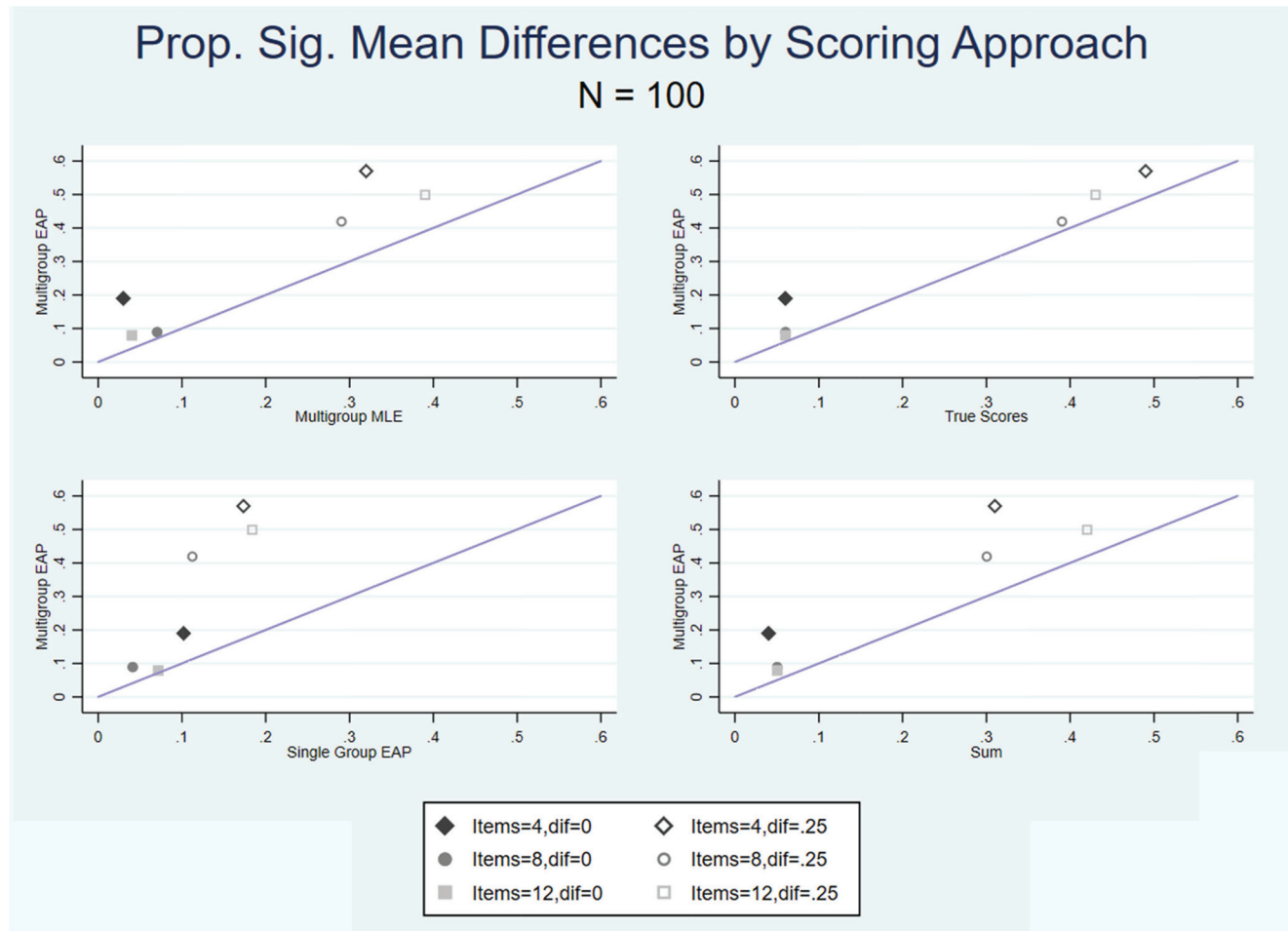
Unidimensional Model. As before, one could simply use a unidimensional IRT model despite having multiple groups and timepoints. However, doing so is not so straightforward. For example, does one simply ignore all timepoints except Time 1? Use all timepoints, but do not model them as if they are distinct? One could fit a unidimensional

model in several ways, including estimating item parameters based on (a) the first timepoint only, (b) item responses from both timepoints (pre/post), or (c) randomly selecting a timepoint for each person. For (b), one could stack the item responses in one long file that ignores the fact that study participants show up multiple times in the data, and calibrate/score those responses. The data could then be reshaped wide with two scores per person to estimate treatment effects. Some limited research suggests that using only the first timepoint versus stacking all the timepoints does not have huge implications for use of scores in growth models (Kuhfeld & Soland, 2020). While these unidimensional approaches may improve on sum scores by allowing, say, items to have different slope parameters, they nonetheless have important limitations, including that they do not account for covariances in scores over time.

Multidimensional Model, Calibrate With Data From Both Timepoints. Alternatively, one could use a MIRT model to estimate latent change before and after treatment (Figure 2, Panel C). Item

Figure 5

Plots of the Proportion of Significant Treatment Effect Estimates for the Two Group Single Timepoint Simulation



Note. See the online article for the color version of this figure.

response data from each timepoint are combined and calibrated simultaneously across the two timepoints. Items are calibrated such that surveys administered before and after the intervention have their own latent variable estimates. Thus, unlike the simple IRT model, the MIRT approach explicitly accounts for differing latent means and variances by timepoint, as well as overtime correlations in the model. Such a model could also be adapted to account for the fact that the same item is administered multiple times during data collection (e.g., by including method factors).

Longitudinal Multigroup IRT Model. One could further expand the MIRT model to fit a longitudinal multigroup MIRT model (LM-MIRT), with differing parameter estimates for control versus treatment groups and by timepoint (see Figure 2, Panel D). Such a model would allow one to relax assumptions including measurement invariance before and after treatment, measurement invariance across groups, and equality of the latent means and variances across groups. In many ways, this calibration approach provides the most flexibility and, perhaps, best matches the nature of data from RCTs. For example, such a model would reflect that pre/post scores can have different covariances between control and treatment groups, as well as

different variances at one or both timepoints and by group. This model would also explicitly include different latent means pre/post treatment for both groups.

Additionally, as Cai, Choi, and Kuhfeld (2016) note, it is implausible in an RCT context to assume full exchangeability of individuals across treatment and control conditions (e.g., the posttreatment latent variables in the control condition cannot be exchanged without consequence to those in the treatment condition). However, standard EAP approaches that do not account for treatment assignment do assume full exchangeability holds. Furthermore, research already demonstrates models like the LM-MIRT tend to do the best job of recovering true treatment effects from pre/post intervention designs (Gorter et al., 2016; Kuhfeld & Soland, 2020; Soland, 2021).

Stage III. Parameter Estimation Decisions

As discussed in Example 1, parameter uncertainty and sample size are factors when fitting an IRT model. Further, the number of parameters typically increases in tandem with the model complexity and number of items in the measure. For instance, the LM-MIRT has many more parameters than the IRT model in Example 1 because the former allows many parameters to vary by group

and timepoint. There is already evidence that sample size affects accurate estimation of item parameters and subsequent IRT score estimates (e.g., Sahin & Anil, 2017; Yang et al., 2012). In short, the measurement model itself could be underpowered, and therefore introduce both noise and bias into treatment effect estimates (though, as we show in the simulations, using a measurement model is often preferable even at relatively small sample sizes).

Stage IV. Scoring Decisions

MLE scoring is usually not recommended for use with multidimensional models. The reasons include that is that MLE scoring does not use information from the population distribution, there can be convergence issues in score estimation, and standard errors can be undefined (Vector Psychometric Group, 2021). Therefore, EAP and MAP scoring approaches are the most widely used for producing scores from MIRT models, including the LM-MIRT. Note that, as in Example 2, the effect of shrinkage will differ substantively across measurement models. A unidimensional model will shrink to a single mean, a MIRT model using both timepoints would shrink to two time-specific means, and the LM-MIRT would shrink toward group- and time-specific means. Arguably, only the LM-MIRT matches the data generating process in a pre/post RCT design where control and treatment groups are manipulated to show differences on the construct over time.

Simulation Study 3

Data Generation and Scoring

In the third simulation study, data were generated using an LM-MIRT model as the true data-generating mechanism. The sample size and survey length conditions were equivalent to the prior example. For this data generation approach, true ability estimates were generated for group g [$g = 0$ for control, $g = 1$ for treatment] and timepoint t [$t = 1, 2$] θ_{gt} . That is, $t = 1$ represents the pretest, and $t = 2$ the posttest. Thus, there were four latent means of interest: μ_{01} (control group pretest), μ_{02} (control group posttest), μ_{11} (treatment group pretest), and μ_{12} (treatment group posttest) where μ_{gt} is the true mean of θ_{gt} . Those true latent means were set equal to $\mu_{01} = 0$, $\mu_{02} = .1$, and $\mu_{11} = 0$. Meanwhile, the posttest score for the treatment group, μ_{12} , was varied such that the mean difference in the gains between the groups equaled 0, .1, and .25.

Further, these data were generated with the following true covariance structure for the control group:

$$\begin{pmatrix} \sigma_{01}^2 & \\ \sigma_{01,02} & \sigma_{02}^2 \end{pmatrix} = \begin{pmatrix} 1 & \\ .9 & 1.4 \end{pmatrix} \quad (1)$$

For the treatment group, the true variance-covariance matrix was:

$$\begin{pmatrix} \sigma_{11}^2 & \\ \sigma_{11,12} & \sigma_{12}^2 \end{pmatrix} = \begin{pmatrix} 1 & \\ .7 & 1.2 \end{pmatrix} \quad (2)$$

These covariances were based on empirical RCT data⁹ and make some important assumptions. For one, the latent variance for the posttreatment is lower for the treatment group. For another, the

covariances between Time 1 and Time 2 are slightly lower for the treatment group. These latent means, variances, and covariances were based on data obtained from an actual RCT (Henry et al., 2012).

Item responses were calibrated and scored using several different measurement models (outlined in Table 1). First, we used a unidimensional IRT model calibrated based on Time 1 (both groups), then those item parameters were used to score all item responses at Time 2. Second, we used a MIRT model that allowed for different means and variances by timepoint, but not by group. Third, we used the LM-MIRT, which allowed means and variances to differ by group and timepoint (both the MIRT and LM-MIRT also account for covariances in scores over time, but not differentially by group for the MIRT model). We then produced sum scores and, as a point of comparison, examined true scores produced during the data simulation. The LM-MIRT models were estimated via the MH-RM algorithm implemented in flexMIRT. Estimates of the person-level scores were produced based on the calibrated item parameters using the EAP scoring approach given the multiple issues associated with MLE scoring when fitting multidimensional IRT models (for consistency, all scores involving a measurement model were produced using EAP).

Estimating Treatment Effects, Type II Errors

Finally, scores were used to estimate treatment effects. In addition to using the scores produced using IRT models, true scores from the data generating process were used to estimate treatment effects, as were sum scores. Treatment effects were produced by taking scores and shaping them long such that the data matrix consisted of columns for the score, group membership, and timepoint. Treatment effect were then estimated by regressing the score on group, timepoint, and a group-by-timepoint interaction, with the coefficient on the interaction representing the estimate of interest. These estimates of the treatment effect were then used to quantify the proportion of significant treatment effect estimates ($\alpha = .05$)¹⁰ across replications (1 – prop. Type II errors).

Results

Table 6 shows the regression-based estimated treatment effect, as well as the standard error on the treatment effect and proportion of significant estimates across the replications. Results for a true treatment effect of .1 *SD*, as well as sample sizes of 150 and 250, are not reported for parsimony. This table helps draw several conclusions about the effect of measurement on treatment effect estimates and Type II errors, all of which are consistent with the results from Study 2. First, only the LM-MIRT scores consistently reproduced the true treatment effects across conditions (other than when using true scores). In particular, bias was substantial for the unidimensional and MIRT

⁹ Henry, D. B., Tolan, P. H., Gorman-Smith, D., & Schoeny, M. E. (2012). Risk and direct protective factors for youth violence: Results from the Centers for Disease Control and Prevention's Multisite Violence Prevention Project. *American Journal of Preventive Medicine*, 43(2), S67–S75.

¹⁰ Results were not sensitive to use of a different significance threshold in terms of substantive conclusions drawn.

Table 6*Results From the Simulation Using Two Groups and Two Timepoints*

		Est. mean diff. (regression)					Group coeff. SEs					Prop. sig.				
N	Items	LM-MIRT	Unidim.	MIRT	Sum	True score	LM-MIRT	Unidim.	MIRT	Sum	True score	LM-MIRT	Unidim.	MIRT	Sum	True score
Mean dif = 0												Prop. Type I errors				
100	4	-0.012	0.009	0.013	0.004	-0.008	0.190	0.198	0.215	0.183	0.214	0.052	0.000	0.000	0.000	0.000
200	4	-0.006	-0.011	-0.006	0.010	-0.002	0.131	0.140	0.152	0.130	0.153	0.083	0.000	0.000	0.021	0.000
500	4	-0.015	-0.004	-0.002	0.001	-0.005	0.083	0.089	0.096	0.082	0.096	0.042	0.000	0.020	0.000	0.000
1,000	4	-0.005	0.005	0.004	0.007	0.001	0.059	0.062	0.068	0.058	0.068	0.104	0.010	0.000	0.021	0.000
100	8	-0.002	0.022	0.017	0.010	0.010	0.202	0.224	0.234	0.175	0.213	0.010	0.000	0.000	0.010	0.000
200	8	0.004	0.015	0.013	0.017	0.004	0.144	0.158	0.165	0.125	0.152	0.000	0.000	0.000	0.000	0.000
500	8	-0.003	0.010	0.008	0.015	-0.004	0.090	0.100	0.104	0.079	0.096	0.021	0.000	0.000	0.021	0.000
1,000	8	0.003	0.007	0.005	0.017	0.004	0.064	0.071	0.073	0.056	0.068	0.000	0.000	0.000	0.000	0.000
100	12	0.003	0.022	0.020	0.022	0.012	0.206	0.229	0.237	0.171	0.215	0.000	0.000	0.000	0.000	0.000
200	12	0.001	0.018	0.014	0.008	0.004	0.144	0.162	0.167	0.121	0.151	0.010	0.000	0.000	0.010	0.000
500	12	0.003	0.002	0.002	0.013	0.000	0.092	0.103	0.105	0.076	0.096	0.010	0.000	0.000	0.000	0.000
1,000	12	-0.003	0.002	0.002	0.013	0.000	0.065	0.073	0.075	0.054	0.068	0.000	0.000	0.000	0.000	0.000
Mean dif = .25												1 - Prop. Type II errors				
100	4	0.251	0.119	0.092	0.171	0.242	0.190	0.119	0.216	0.182	0.214	0.250	0.020	0.000	0.125	0.083
200	4	0.244	0.098	0.069	0.179	0.248	0.131	0.098	0.152	0.129	0.153	0.479	0.030	0.000	0.229	0.250
500	4	0.235	0.105	0.075	0.168	0.245	0.083	0.105	0.096	0.082	0.096	0.792	0.141	0.010	0.531	0.833
1,000	4	0.245	0.113	0.080	0.173	0.251	0.059	0.062	0.068	0.057	0.068	0.969	0.485	0.030	0.896	0.990
100	8	0.252	0.159	0.132	0.188	0.260	0.202	0.159	0.234	0.173	0.213	0.146	0.010	0.000	0.073	0.063
200	8	0.256	0.152	0.128	0.195	0.254	0.144	0.152	0.165	0.123	0.152	0.406	0.030	0.010	0.323	0.333
500	8	0.248	0.148	0.122	0.193	0.246	0.090	0.148	0.104	0.078	0.096	0.823	0.242	0.081	0.750	0.854
1,000	8	0.255	0.145	0.120	0.196	0.254	0.063	0.070	0.073	0.055	0.068	1.000	0.576	0.172	1.000	1.000
100	12	0.255	0.165	0.146	0.199	0.262	0.205	0.165	0.237	0.169	0.215	0.135	0.000	0.000	0.135	0.073
200	12	0.249	0.162	0.139	0.186	0.254	0.144	0.162	0.167	0.120	0.151	0.354	0.051	0.020	0.292	0.313
500	12	0.254	0.146	0.127	0.191	0.250	0.092	0.146	0.106	0.076	0.096	0.885	0.192	0.071	0.792	0.896
1,000	12	0.247	0.146	0.126	0.191	0.250	0.065	0.073	0.075	0.053	0.068	0.990	0.576	0.313	1.000	1.000

Note. LM-MIRT = longitudinal multigroup - multidimensional item response theory.

models. To help make this point clearer, Figure 6 shows density plots of estimated treatment effects by measurement model for a condition with 12 items, sample size of 500 per group, and a true treatment effect of .25 SDs. As the figure shows, the LM-MIRT is the only approach that produces scores that recover the true treatment effect consistently, and the other IRT-based approaches often understate the true treatment effect by approximately half.

Second, due to shrinkage, the LM-MIRT scores are far more likely to produce Type I errors than the other scoring methods, especially for a four-item scale. For example, when the true difference in the gains by group equals zero in the data-generating model and the survey consists of four items, significant treatment effects were found anyway, in some cases nearly 10% of the time ($N = 200$). Thus, while the LM-MIRT does the best job of recovering true treatment effects, it can produce spurious results (Type I errors) when the survey scales are very short and respondent sample size is small. At the same time, Type I errors were quite low even for EAP using the LM-MIRT when the survey scales were 8 items or longer. In short, the implications of the measurement model and scoring approach differ considerably for very short scales versus those with at least eight items.

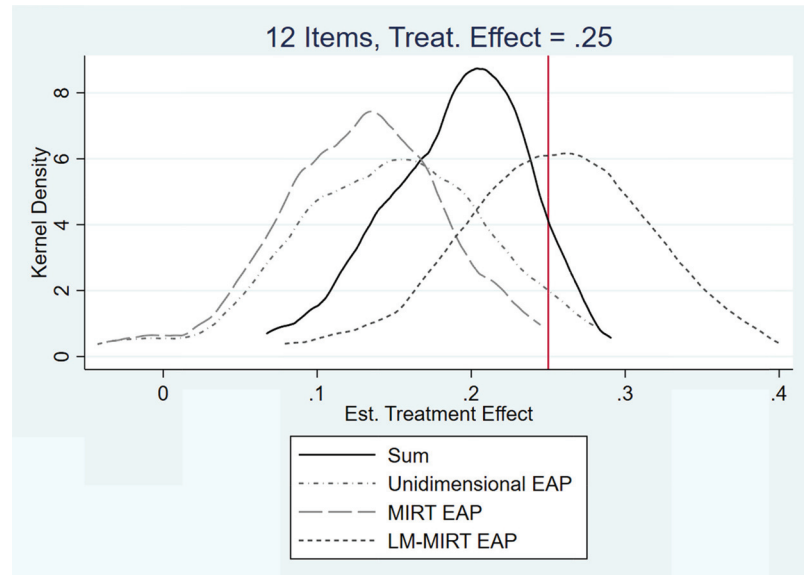
Third and conversely, when the true treatment effect was .25 SDs, the LM-MIRT scoring approach was much more likely to avoid Type II errors. If one were to take one minus the proportion of significant results for the condition with a treatment effect of .25

SDs, then that number would be the Type II error rate. For example, with a sample size of 100 per group and only four items, 75% of replications produced a Type II error for the LM-MIRT model compared with 88% for sum scores. In some cases, with small sample sizes and a four-item survey scale, LM-MIRT scores produced twice as many significant results as when using other scores.

To help make these tradeoffs clearer, Figure 7 shows bar plots of the proportion of significant effects by measurement model for conditions with four versus 12 items, 100 versus 500 respondents per group, and a true treatment effect of .25 SDs. As the bottom row with 500 respondents shows, when the sample size is a bit larger, the LM-MIRT is far superior to other measurement models, producing more significant results, and doing so in a way that mirrors results using true scores. By contrast, the first row shows results for a sample size of 100. While the LM-MIRT once again outperforms the other models, it also drastically outperforms the true scores, especially with only four items. Though such an outcome might seem desirable, a proportion of those significant results are actually negative, which means one would wrongly conclude the treatment was not only ineffective, but actually had a detrimental impact. Further, assuming one did not know the true treatment effect (as we do here), outperforming the true score means we are getting many more significant results than power analyses would suggest, raising the specter of Type I errors and other spurious conclusions.

Fourth, these differences almost uniformly resulted because the standard errors on the treatment effect estimate were smaller, on average, for LM-MIRT scores than the true scores. This likely

Figure 6
Density Plots of Treatment Effect Estimates for the Multigroup Multi-Time-Point Simulation



Note. LM-MIRT = longitudinal multigroup - multidimensional item response theory; EAP = expected a posteriori. See the online article for the color version of this figure.

occurred because the LM-MIRT scores were shrunk toward group- and time-specific means. Thus, when scales were less reliable and introduced more uncertainty, scores tended to be shrunk toward those means to a greater degree. While similar shrinkage occurred for the other IRT models in this simulation due to EAP scoring, the unidimensional model shrunk scores to a single mean, and the MIRT model to two means (by contrast, the LM-MIRT shrunk scores toward four means, treatment and control pre- and post-treatment). Thus, the bias produced by those other IRT models when they shrank scores to the wrong means was much more impactful than their standard errors. Meanwhile, LM-MIRT scores also reduced Type II errors relative to sum scores (likely due in part to sum scoring wrongly constraining the slope parameters equal across items; Soland, 2021).

Empirical Study

To help show that the effects of the various measurement decisions we consider here are not purely a function of how we simulated data, we also used item response data from a largescale RCT that examined the impact of the GREAT Families Program on student aggressive behavior. The sample included over 5,000 students in middle school grades. The initial study found a positive and significant effect of the intervention on school norms for aggression, an outcome of interest (Henry et al., 2012; Simon et al., 2009). School norms for aggression were assessed using the Norms for Aggression and Alternatives scale (Dymnicki et al., 2011). In parallel with the simulation studies, scores on the School Norms for Aggression scale were produced using a unidimensional IRT model calibrated at preintervention using the control group only

and the LM-MIRT model. In addition, sum scores standardized relative to the mean and variance of the control group preintervention were used. Treatment effects were estimated for these three sets of scores using a multilevel model that regressed postintervention scores on preintervention scores and treatment status, and included a site random intercept.

Additional information on the study and detailed results are provided in the online supplemental materials. Not unlike in the simulation studies, using a standardized sum score produced a treatment effect estimate that was ~20% lower than when using the LM-MIRT model.¹¹ Further, the sum score approach closely matches the treatment effect reported in Simon et al. (2009), which may indicate that the reported treatment effect understated the true treatment effect. Also similar to the simulation studies, the other IRT-based approaches tended to produce treatment effects that were lower than those using sum scores and an LM-MIRT model. Most relevant to the simulation studies, the confidence interval was smallest for the LM-MIRT given it produced smaller standard errors on the treatment effect than using the other scores. In short, results from the empirical study are consistent with those from the simulation studies.

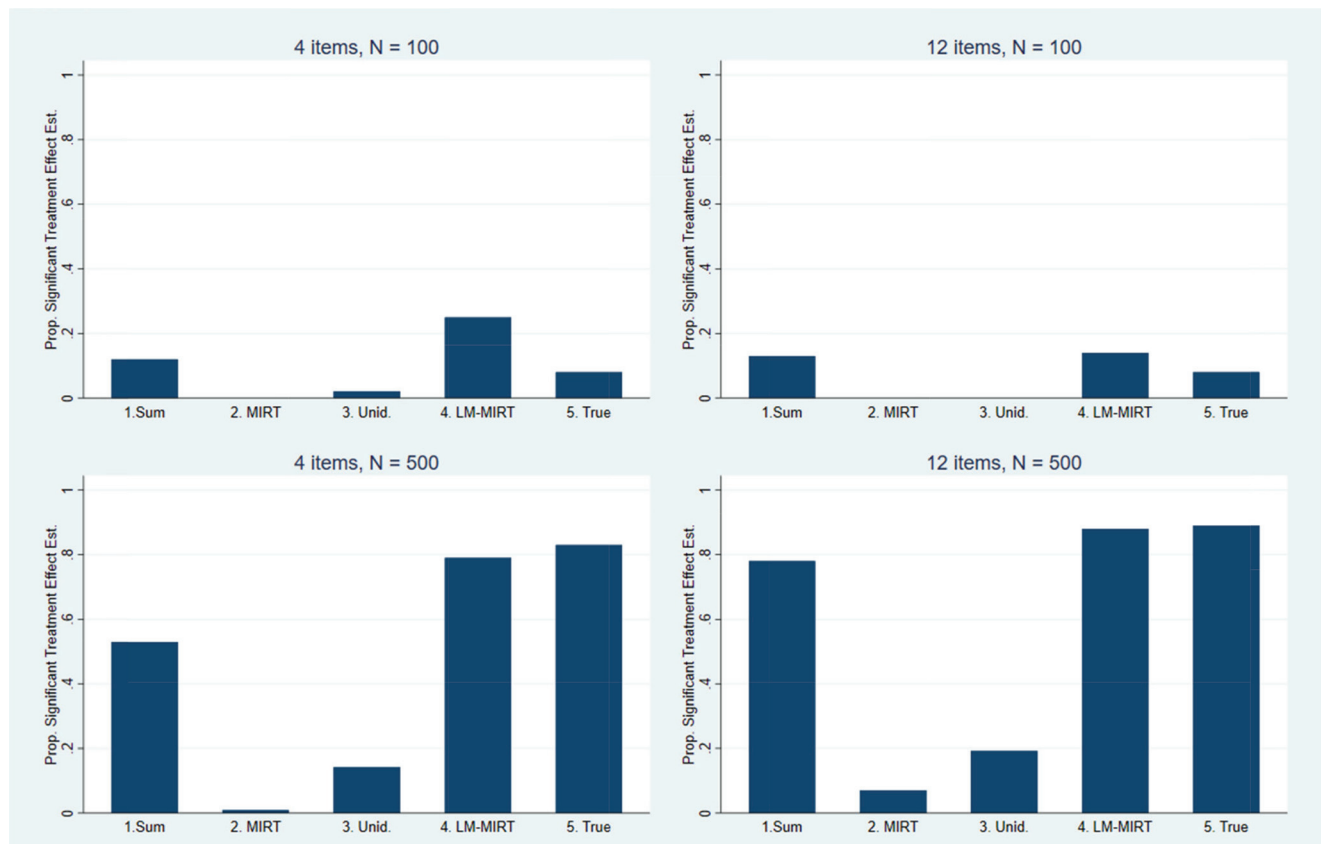
Discussion and Implications for Psychological Research

Much effort is often put into designing psychological studies. Yet, very little consideration is given to how decisions about

¹¹ Note that standardizing a sum score is not sufficient to mitigate scaling differences between sum and IRT scores, which likely accounts for some of these differences.

Figure 7

Bar Plots Showing the Proportion of Significant Treatment Effect Estimates by Scoring Approach, True Effect = .25 SDs



Note. IRT = item response theory; LM-MIRT = longitudinal multigroup - multidimensional item response theory; MLE = maximum likelihood estimation; EAP = expected a posteriori; RCT = randomized control trial. See the online article for the color version of this figure.

scoring survey scales might affect the outcome of these RCTs. In fact, most such RCTs in psychology and education use sum scores despite the untenable assumptions they often make (McNeish & Wolf, 2020). In our own study, we used simulation and empirical evidence to examine how measurement decisions affect three common study designs/estimands in psychology. We examined that impact (especially related to measurement model and scoring phases) on recovery of true means and variances (including group differences), as well as on Type I and II error rates. The intent was to help researchers understand the implications of such decisions for their own work, as well as to help the broader field better understand how measurement affects replication across studies. For the primary calibration and scoring approaches discussed, we provide path diagrams, equations, and code (flexMIRT and R) that can be used to replicate our results in the online supplemental materials. Our analyses produced several results that can help guide decisions made by researchers, especially conducting RCTs.

Considerations for Researchers by Stage

Given our findings, what is a conscientious researcher using survey scales to do? Table 7 presents considerations (and, in many cases, potential pitfalls) for each stage of the measurement process

shown in Figure 1 by example/study. We discuss those considerations below and, where defensible, provide guidance.

Stage I. Whether to Use a Measurement Model?

In general, results suggest researchers should probably use a measurement model rather than sum scores. One exception is when the sample size is very small such that item parameters in a measurement model cannot be properly estimated, or such that using a more complex measurement model (which often necessitates EAP scoring) overly shrinks scores and produces Type I errors. However, in most cases with a sufficient sample size (roughly 200 or more respondents for short measures, likely larger sample sizes for longer measures), a measurement model is almost certainly justified. While the assumptions of sum scores can occasionally be met, they are quite strong and often violated. Such violations typically induce bias into estimated means and mean differences, a result found here and in many other studies (e.g., McNeish & Wolf, 2020).

Figure 7, which examines simulated studies akin to a pre/post RCT with a true treatment effect of .25 SDs, brings this tradeoff into specific relief. When the sample size is only 100, multigroup multi-time-point IRT models tend to produce far more significant results than when using true scores, raising the specter of Type I errors. However, with a

Table 7*Considerations for Researchers Trying to Score Survey Items*

Study example/ sim. study Groups, timepoints	Considerations							
	Example 1		Example 2		Presented elsewhere		Example 3	
	Groups = 1	Time = 1	Groups = 2	Time = 1	Groups = 1	Time = 2+	Groups = 2	Time = 2
Hypothetical study design	Est. means		Est. treat/control diff., sub-group diff./gaps		Est. growth		RCT with pre/post design	
Stage 1. Measurement model?	Using sum scores can misweight items in ways that bias means and variance estimates		Same as Example 1, with mean difference estimates biased		Same as Example 1, with growth estimates (means/variances) biased		Same as Example 1, with pre/post estimates biased	
Stage 2. Measurement model	Not many competing options (unidimensional IRT model sufficient)		Using an IRT model that does not match the study design (here, multigroup IRT) can introduce bias		Using an IRT model that does not match the study design (here, longitudinal MIRT) can introduce bias		Using an IRT model that does not match the study design (here, LM-MIRT) can introduce bias	
Stage 3. Calibration	Sample sizes larger than 200 individuals needed for unbiased parameter estimates		Sample sizes greater or equal to 200 individuals (per group) needed for unbiased parameter estimates		Not examined in this study		Sample sizes greater or equal to 200 individuals (per group) needed for unbiased parameter estimates	
Stage 4. Scoring approach	MLE (max and miss) can introduce bias; EAP unbiased		Same as Example 1, bias most pronounced for four-item scale		Not examined		Not examined; EAP only feasible option for more complex MIRT models	
Bias	MLE-produced bias can affect error rates; EAP shrinkage increased Type I errors, reduces Type II		Same as Example 1, Type I errors especially high for EAP with four- to eight-item scales		Not examined		For EAP scoring, Type I error rates still high when using four-item scale, but not eight- to 12-item scale	
Type I/II error rates								

Note. IRT = item response theory; MIRT = multidimensional item response theory; MLE = maximum likelihood estimation; EAP = expected a posteriori; RCT = randomized control trial. Presented elsewhere = This scenario was not presented in the current study, but was examined in depth by Kuhfeld and Soland (2020) and Bauer and Curran (2016). Larger sample sizes may be needed for accurate calibration for longer measures than those examined in this study or for datasets with a significant proportion of missing data.

sample size of 500, the multigroup multi-time-point IRT model substantively outperforms sum scores. For example, when the sample size is 500 and there are only four items, the LM-MIRT produces significant results in line with true scores, yet produces significant results across replications at a rate nearly thirty percentage points higher compared to sum scores. In short, with a reasonable sample size, the LM-MIRT tends to avoid Type I errors while drastically reducing Type II errors compared with sum scores.

Further, the assumptions of sum scores become greater in number and typically less tenable as the number of groups and timepoints increase. They can introduce particularly large bias when examining growth over time (e.g., Kuhfeld & Soland, 2020), and when using a pre/post RCT design (e.g., Soland, 2021). Though slope estimates from latent growth curve models were not examined in the current study, Kuhfeld and Soland (2020) found that using sum scores compared to LM-MIRT scores resulted in severe understatement of latent slopes means and the variability of those slopes. In some cases, mean growth was underestimated by nearly half when using sum scores.

Stage II. Choosing a Measurement Model

Our results also indicate that one of the best ways to avoid bias may be to select a measurement model that matches the nature of the study design (and, therefore, the data-generating process). Essentially, this recommendation amounts to selecting models that reflect the number

of groups and timepoints in the study design. For example, if one has a multi-time-point model, but uses an IRT model that assumes there is only one timepoint, considerable bias can be introduced into estimates of growth, as well as pre/post treatment contrasts. At the same time, researchers should always look for evidence of model misspecification, even when using an IRT model that ostensibly matches the study design. For example, research shows that using a more complex IRT model can actually introduce more bias than when using sum scores if misspecified (Rhemtulla et al., 2020).

While there are many forms that such misspecification can take, our results indicate that one is far more likely to end up with model misspecification by using a sum score or IRT model that does not match the nature of the research question of interest than when using an IRT model that does. For example, measurement model misspecification would be likelier, in our view, if one assumes that scores from a longitudinal study are not correlated over time (as one would when using a sum score or unidimensional IRT model in a growth study), or if one compounds such an assumption with assuming that there is only a single group (as one would do fitting any model we considered except the LM-MIRT to an RCT with a pre/post design). Further, these misspecifications seem even less likely to occur in experimental settings, where the design involves an intervention that differentially impacts groups, including over time. Nonetheless, these decisions are impacted by the scoring approach used in Stage III and, in particular, whether shrinkage is employed.

Finally, one should consider the intended score uses when making these decisions. We have focused in our study on cases where researchers are using scores from surveys to estimate group-level inferences, such as mean differences within and across timepoints. In this scenario, our recommendation of including important conditioning information in the calibration/scoring model is consistent with decades of work done in the large-scale assessment area (Mislevy et al., 1992; von Davier & Sinharay, 2013), which has highlighted potential biases that result from not including conditioning information in the population model. By comparison, in the scenario where scores are being used to make individual-level decisions, there could be unintended consequences of the inclusion of conditioning information in scoring. Purely hypothetically, if boys traditionally score lower overall than girls on a self-efficacy measure, a boy could receive a lower EAP score from a multigroup IRT model (where the population mean for the boys group is estimated to be lower than the girls' distribution) than a girl with the same pattern of correct responses. If scores are tied to, say, an intervention, this scenario raises all kinds of fairness concerns, which is why most educational assessments are currently scored using MLE (where no information about the population is included).

Stage III. Choosing an Estimation Approach

While item parameter estimation was not a focal point of our study, we did examine the recovery of item parameters in our various motivating examples to ensure any score recovery issues were not being driven by poorly estimated item parameters. Our results (see Tables S3.1–S3.3 in the online supplemental materials) indicated that standard IRT estimation approaches implemented in flexMIRT worked well (even for forms as short as four items) when there were greater than 200 students in the sample. While others have done deeper investigations into minimal sample sizes needed for accurate estimation of IRT parameters across a range of conditions (e.g., Sahin & Anil, 2017), our findings further reinforce the point that IRT calibration and scoring should probably not be attempted for studies with less than 200 participants. However, if a measure developer has precalibrated the measure with an appropriate reference population, the size of a researcher's study sample is then less consequential and scoring can be conducted with the precalibrated item parameters.

Stage IV. Choosing a Scoring Approach

Deciding between MLE and EAP scoring involves several complex tradeoffs. First, when considering MLE, one must remember that scores cannot easily be produced when the respondent uses only a single response category. This scenario is particularly likely for very short survey scales because selecting, say, *strongly agree* for all four items is plausible for respondents providing accurate and truthful responses. As our results show, replacing such undefined scores with an arbitrary maximum (in our case the software default) can introduce considerable bias into estimated means, as can simply treating those scores as missing (depending on the missingness patterns). As for variances, MLE often tends to inflate SDs of scores, which reduces Type I errors, but also substantively increases Type II errors relative to when using true scores. Finally, MLE often cannot be used in conjunction with more complex IRT models, and is therefore not feasible for a model like the LM-MIRT.

By comparison, EAP scoring has its own advantages and disadvantages, and they are often the converse of those when using MLE. In particular, EAP tends to produce very little bias in estimated means, regardless of the study design. Further, it can be used even for complex IRT models like the LM-MIRT. Another advantage of EAP estimation is that it is a noniterative approach which saves processing time, avoids local minima/maxima, and provides for a fixed set of equations for computing each examinee's score.

However, due to shrinkage, EAP scoring can downwardly bias the variance of scores, especially when scales are short (e.g., four items) and sample sizes are small. These biased variances, in turn, reduce Type II errors, but can increase Type I errors, in some cases hugely. For example, when comparing means for two groups at a single timepoint, the Type I error rate is often well above the α rate of .05 and, while using 12 items makes the Type I error rate fairly comparable to when using sum scores, it remains higher. In the case of two groups and two timepoints, Type I errors are quite low when the survey scale has eight or more items, but is very high when there are only four items. These scoring issues provide yet another argument for developing survey scales that are of a decent length (ideally 12 or more items).

Turning to Type II errors, EAP does the best job of avoiding them. Yet, in some cases it does too well, producing Type II error rates that are better than when using true scores. To further enumerate why this result is less than ideal, our findings indicate that using EAP scoring is more likely to produce significant results the less reliable one's survey scale. That is, if all one cared about was avoiding Type II errors, then the best study design (purely from a measurement perspective) might involve an unreliable scale and a very small sample size. These findings raise serious concerns in the context of the replication "crisis" in psychology. Some researchers may be finding spuriously significant results simply by using EAP scoring with very short survey scales. Conversely, researchers using this approach may be getting far lower Type II error rates compared to others who have simply elected to use a different scoring approach.

In general, given bias and Type I/II error rates, EAP may not be justifiable for short survey scales and with small sample sizes. Concomitantly, given similar considerations, EAP may be the preferred approach for longer survey scales (12+ items) and larger sample sizes. For example, with longer scales, EAP reduces Type II errors and has Type I error rates that are higher than when using true scores, but not dramatically so. Ultimately, the best approach may involve producing scores using EAP and another approach like MLE for the purposes of consistency. When a scale is even longer (e.g., including 20 items), there should be very little scale shrinkage and therefore likely a minimal difference between MLE and EAP scores.

Broader Research Implications

Our findings also suggest that research bodies tasked with disseminating best practices in program evaluation should probably pay more attention to measurement. For example, in education, the What Works Clearinghouse disseminates best practices for experimental and quasi-experimental designs. Whereas copious research from the Clearinghouse considers benefits and limitations of various quasi-experimental models, the current version of their Standards Handbook mentions reliability, but generally does not consider other measurement issues, and makes no mention of IRT. Given results from this

study and work by others (Bauer & Curran, 2016; Gortner et al., 2016; Kuhfeld & Soland, 2020; Soland, 2021), there is a strong argument that best practices for evaluating programs should include, at minimum, reporting the type of measurement model used and the calibrated item parameters so they can be used in future studies.

Limitations

A few limitations of this study bear mention. First, we did not examine a condition in which items were precalibrated on a larger sample. We made that choice because studies using short survey scales in psychology so rarely provide calibrated item parameters that other researchers can use (e.g., Flake et al., 2017). Further, precalibration is less likely to have been conducted for items in a multigroup calibration model, which is the context of two of our three motivating examples.

Second, we do not examine the effect of clustering of respondents on Type I and II errors and its interaction with scoring approach. Educational and psychological assessment is often conducted in settings where students are organized in nested groups, such as schools, clinics, or classrooms. Ignoring this nesting can have nontrivial impacts on the estimation of measurement models (Zyphur et al., 2008). Oftentimes, cluster RCTs are underpowered relative to RCTs in which randomization occurs at the person level (e.g., Heo & Leon, 2008). One cannot be sure how clustering would impact results from this study.

Conclusion

While much attention is given to aspects of study design in psychology, including RCTs, relatively little is given to how calibration and scoring decisions affect estimates means and variances, including the Type I and Type II error rates impacted by those parameters. Our results suggest this oversight could have important implications for the findings in psychological studies, including both individual RCTs and replications across studies. We detailed three examples, each representing a common study design in psychology, then simulated data according to each to show the magnitude of how measurement decisions impact results. Finally, we articulated tradeoffs in various measurement decisions to support researchers attempting to conduct studies using survey scales.

In general, our results suggest that the decisions that go into scoring a survey can affect the bias in one's estimated means and SDs, both of which are typically used in hypothesis testing. And, while differences in that bias across our results make it difficult to give a simple overarching recommendation, we do tend to find that using a measurement model that best matches the ostensible data-generating process introduces the least bias (e.g., a multigroup IRT model for a control-treatment contrast). To help match the model to the data-generating process, we provide a scoring decision flowchart in Figure 1 that can help researchers better identify such models in the context of a particular study design. As for specifics, using an IRT model is preferable to employing sum scores with a sufficient sample size to calibrate item parameters (roughly 200 or more respondents). In terms of scoring, MLE tends to produce biased means under many plausible scenarios, and both MLE and EAP often produce biased variances. When a survey scale is sufficiently long (at least 8 items, and ideally 12 or more), EAP scoring may be preferable because it can be employed in conjunction with multidimensional IRT models and reduces

Type II errors. However, when survey scales are short and sample sizes are small, EAP scoring can overly shrink scores to the mean such that Type I errors are unacceptably high.

References

- Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement*, 16(1), 87–96.
- Bauer, D. J., & Curran, P. J. (2016). The discrepancy between measurement and modeling in longitudinal data analysis. In J. R. Harring, L. M. Stapleton & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research* (pp. 3–38). Information Age Publishing.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4), 443–459. <https://doi.org/10.1007/BF02293801>
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4), 431–444.
- Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35(3), 307–335. <https://doi.org/10.3102/1076998609353115>
- Cai, L. (2017). flexMIRT version 3.51: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Vector Psychometric Group.
- Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item response theory. *Annual Review of Statistics and Its Application*, 3(1), 297–321. <https://doi.org/10.1146/annurev-statistics-041715-033702>
- Cai, L., Choi, K., & Kuhfeld, M. (2016). On the role of multilevel item response models in multi-site evaluation studies for serious games. In H. F. O'Neil, E. L. Baker, & R. Perez (Eds.), *Using games and simulations for teaching and assessment* (pp. 280–301). Taylor & Francis.
- Curran, P. J., Cole, V. T., Bauer, D. J., Rothenberg, W. A., & Hussong, A. M. (2018). Recovering predictor–criterion relations using covariate-informed factor score estimates. *Structural Equation Modeling*, 25(6), 860–875. <https://doi.org/10.1080/10705511.2018.1473773>
- Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, 76(5), 741–770. <https://doi.org/10.1177/0013164415607618>
- Dymnicki, A. B., Antônio, T., & Henry, D. B. (2011). Levels and growth of specific and general norms for nonviolence among middle school students. *Journal of Adolescence*, 34(5), 965–976. <https://doi.org/10.1016/j.adolescence.2010.11.012>
- Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, 75(3), 474–497. <https://doi.org/10.1007/s11336-010-9161-9>
- Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58(3), 357–381. <https://doi.org/10.1177/0013164498058003001>
- Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 3(4), 456–465.
- Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological & Personality Science*, 8(4), 370–378. <https://doi.org/10.1177/1948550617693063>
- Gehlbach, H., & Hough, H. J. (2018). *Measuring Social Emotional Learning through Student Surveys in the CORE Districts: A Pragmatic Approach to Validity and Reliability*. Policy Analysis for California Education, PACE. <https://eric.ed.gov/?id=ED591082>
- Gortner, R., Fox, J. P., Apeldoorn, A., & Twisk, J. (2016). Measurement model choice influenced randomized controlled trial results. *Journal of Clinical Epidemiology*, 79, 140–149. <https://doi.org/10.1016/j.jclinepi.2016.06.011>

- Henry, D. B., Cartland, J., Ruchcross, H., & Monahan, K. (2004). A return potential measure of setting norms for aggression. *American Journal of Community Psychology*, 33(3-4), 131-149. <https://doi.org/10.1023/B:AJCP.0000027001.71205.dd>
- Henry, D. B., Tolan, P. H., Gorman-Smith, D., & Schoeny, M. E. (2012). Risk and direct protective factors for youth violence: Results from the Centers for Disease Control and Prevention's Multisite Violence Prevention Project. *American Journal of Preventive Medicine*, 43(2), 567-575. <https://doi.org/10.1016/j.amepre.2012.04.025>
- Heo, M., & Leon, A. C. (2008). Statistical power and sample size requirements for three level hierarchical cluster randomized trials. *Biometrics*, 64(4), 1256-1262. <https://doi.org/10.1111/j.1541-0420.2008.00993.x>
- Kolen, M. J., & Brennan, R. L. (2013). *Test equating: Methods and practices*. Springer Science & Business Media.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. (3rd ed.). Springer-Verlag.
- Kuhfeld, M., & Soland, J. (2020). Avoiding bias from sum scores in growth estimates: An examination of IRT-based approaches to scoring longitudinal survey responses. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000367>
- Lindley, D. V., & Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(1), 1-18.
- McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 1(1), 1-19.
- Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge. <https://doi.org/10.4324/9780203821961>
- Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling Procedures in NAEP. *Journal of Educational Statistics*, 17, 131-154.
- Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, 25(1), 30-45. <https://doi.org/10.1037/met0000220>
- Sahin, A., & Anil, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, 17(1), 321-335.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(S1), 1-97. <https://doi.org/10.1007/BF03372160>
- Schmidt, F. L., & Hunter, J. E. (2015). Meta-analysis of correlations corrected individually for artifacts. *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed., pp. 87-164). SAGE Publications. <https://doi.org/10.4135/9781483398105>
- Simon, T. R., Ikeda, R. M., Smith, E. P., Reese, L. E., Rabiner, D. L., Miller, S., Winn, D.-M., Dodge, K. A., Asher, S. R., Horne, A. M., Orpinas, P., Martin, R., Quinn, W. H., Tolan, P. H., Gorman-Smith, D., Henry, D. B., Gay, F. N., Schoeny, M., Farrell, A. D., . . . Allison, K. W. (2009). The ecological effects of universal and selective violence prevention programs for middle school students: A randomized trial. *Journal of Consulting and Clinical Psychology*, 77(3), 526-542. <https://doi.org/10.1037/a0014395>
- Smith, E. P., Gorman-Smith, D., Quinn, W. H., Rabiner, D. L., Tolan, P. H., & Winn, D. M. (2004). Community-Based multiple family groups to prevent and reduce violent and aggressive behavior: The GREAT Families Program. *American Journal of Preventive Medicine*, 26(1, Suppl), 39-47. <https://doi.org/10.1016/j.amepre.2003.09.018>
- Soland, J. (2021). Evidence That Selecting an Appropriate Item Response Theory-Based Approach to Scoring Surveys Can Help Avoid Biased Treatment Effect Estimates. *Educational and Psychological Measurement*. Advance online publication. <https://doi.org/10.1177/00131644211007551>
- Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, 15(3), e2000797.
- Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73-140). Erlbaum. <https://doi.org/10.4324/9781410604729-8>
- Thissen, D., & Wainer, H. (2001). *Test scoring*. Erlbaum. <https://doi.org/10.4324/9781410604729>
- van der Linden, W. J. (Ed.). (2018). *Handbook of Item Response Theory: Three Volume Set*. CRC Press.
- Vector Psychometric Group. (2021). flexMIRT frequently asked questions. <https://vpgcentral.com/software/flexmirt/support/frequently-asked-questions/>
- von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI Monograph Series*, 2, 9-36.
- von Davier, M., & Sinharay, S. (2013). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. CRC Press.
- Williams, R. H., Zimmerman, D. W., & Zumbo, B. D. (1995). Impact of measurement error on statistical power: Review of an old paradox. *The Journal of Experimental Education*, 63(4), 363-370.
- Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12(1), 58-79. <https://doi.org/10.1037/1082-989X.12.1.58>
- Wolf, R. (2021). *Average differences in effect sizes by outcome measure type*. <https://files.eric.ed.gov/fulltext/ED610568.pdf>
- Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement*, 72(2), 264-290. <https://doi.org/10.1177/0013164411410056>
- Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research*, 81(2), 267-301. <https://doi.org/10.3102/0034654311405999>
- Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (pp. 129-131). American Council on Education and Praeger.
- Zyphur, M., Kaplan, S., & Christian, M. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics*, 12(2), 127-140. <https://doi.org/10.1037/1089-2699.12.2.127>

Received September 24, 2021

Revision received March 16, 2022

Accepted March 29, 2022 ■