The InterModel Vigorish as a lens for understanding (and quantifying) the value of item response modeling

Supplemental Information (SI)

Contents

Code to replicate analysis is available at `https://github.com/intermodel-vigorish`.

S1 Examples

In this section we provide code for computation of the IMV with simulated and real data. The code is also available online.[1]

S1.1 Computing the IMV

The below function will compute the IMV. It requires three core arguments: a vector of responses `resp` and then predictions from two models, `pv1` and `pv2`. This function will get used below in calculations of the IMV in both simulated and empirical examples.

Computing the IMV

```
imv<-function (resp, pv1, pv2, eps = 1e-06)
{
    pv1 <- ifelse(pv1 < eps, eps, pv1)
    pv2 <- ifelse(pv2 < eps, eps, pv2)
    pv1 <- ifelse(pv1 > 1 - eps, 1 - eps, pv1)
    pv2 <- ifelse(pv2 > 1 - eps, 1 - eps, pv2)
    # Log likelihood
    ll <- function(x, p) {
        z <- log(p) * resp + log(1 - p) * (1 - resp)
        z <- sum(z)/length(x)
        exp(z)
    }
    loglik1 <- ll(resp, pv1)
    loglik2 <- ll(resp, pv2)
    getcoins <- function(a) {
        f <- function(p, a) abs(p * log(p) +
                                (1 - p) * log(1 - p) - log(a))
        nlminb(0.5, f, lower = 0.001, upper = 0.999, a = a)$par
    }
    c1 <- getcoins(loglik1)
    c2 <- getcoins(loglik2)
    ew <- function(p1, p0) (p1 - p0)/p0
    imv <- ew(c2, c1)
    imv
}
```

[1] https://github.com/intermodel-vigorish/imv-irt/blob/main/examples/imv.R

S1.2 Simulated Example

We can use the `imv()` function (see SI-S1.1) to compute the IMV for predictions from the 1PL versus the 2PL when the 2PL (with $\sigma = 0.5$) is the data-generating model. Note that the IMV is computed with `resp.test`, a second set of out-of-sample responses generated from the underlying probabilities (the ability and item parameters are estimated based on the 'in-sample' data `resp`).

An example with simulated data

```
##simulate data
set.seed(170301)
N<-10000
ni<-50
th<-rnorm(N)
b<-rnorm(ni)
a<-exp(rnorm(ni,sd=.5))
k<-outer(th,b,'-')
k<-matrix(a,nrow=N,ncol=ni,byrow=TRUE)*k
##estimate 1pl and 2pl
p<-1/(1+exp(-k))
resp<-matrix(rbinom(N*ni,1,p),nrow=N,ncol=ni,byrow=FALSE)
resp.test<-matrix(rbinom(N*ni,1,p),nrow=N,ncol=ni,byrow=FALSE)
resp<-data.frame(resp)
names(resp)<-paste("item",1:ncol(resp))
library(mirt)
m1<-mirt(resp,1,'Rasch')
m2<-mirt(resp,1,'2PL')
p.est<-list()
mods<-list(m1,m2)
##get predictions, compute imv
for (i in 1:length(mods)) {
    m<-mods[[i]]
    th.est<-fscores(m)
    est<-coef(m,simplify=TRUE,IRTpars=TRUE)$items
    k<-outer(th.est[,1],est[,2],'-')
    k<-matrix(est[,1],nrow=N,ncol=ni,byrow=TRUE)*k
    p<-1/(1+exp(-k))
    p.est[[i]]<-p
}
imv(as.numeric(resp.test),
    pv1=as.numeric(p.est[[1]]),
    pv2=as.numeric(p.est[[2]])
    )
```

S1.3 Empirical Example

We also offer an example analysis of empirical data (see Gilbert, Kim, & Miratrix, 2023) drawn from item response data available via the IRW (Domingue & Kanopka, 2023). We again compare predictions from the 1PL and 2PL but this time based on cross-validation. Analysis again uses the `imv()` function (see SI-S1.1).

An example with empirical data

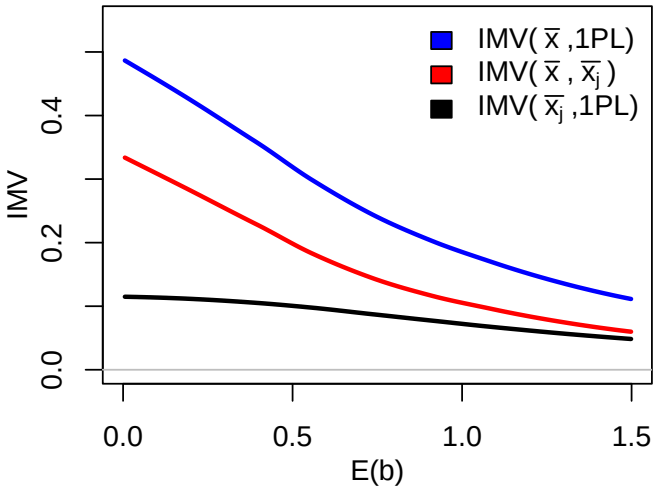
```
##example analysis of one dataset
set.seed(170301)
library(mirt); library(redivis); library(irw)
dataset <- redivis::user("datapages")$
    dataset("item_response_warehouse",version='v2.0')
df <- dataset$table("content_literacy_intervention")$to_data_frame()
df$item<-paste("item_",df$item,sep='')
##cross-validation for models estimated in mirt
ntimes<-4
df$gr<-sample(1:ntimes,nrow(df),replace=TRUE)
omega<-numeric()
for (i in 1:ntimes) {
    x<-df
    x$oos<-ifelse(x$gr==i,1,0)
    x0<-x[x$oos==0,]
    resp0<-data.frame(irw::long2resp(x0))
    id<-resp0$id
    resp0$id<-NULL
    m0<-mirt(resp0,1,'Rasch')
    ni<-ncol(resp0)
    s<-paste("F=1-",ni,"
             PRIOR = (1-",ni,", a1, lnorm, 0.0, 1.0)",sep="")
    model<-mirt.model(s)
    m1<-mirt(resp0,model,itemtype=rep("2PL",ni),
             method="EM",
             technical=list(NCYCLES=10000))
    ##
    z0<-irw::getp(m0,x=x[x$oos==1,],id=id)
    z1<-irw::getp(m1,x=x[x$oos==1,],id=id)
    z0<-z0[,c("item","id","resp","p")]
    names(z0)[4]<-'p1'
    z1<-z1[,c("item","id","p")]
    names(z1)[3]<-'p2'
    z<-merge(z0,z1)
    omega[i]<-imv(z$resp,z$p1,z$p2)
}
mean(omega)
```

S2 Simulation results for dichotomous item response models

S2.1 The IMV as a function of $\mathbb{E}(b)$

We illustrate the behavior of the IMV as a function of the mean difficulty of the measure, $\mathbb{E}(b_j)$ where b_j represents the difficulty parameter for item j (where $c_j = 0$ and $a_j = 1$ in Eqn 6 of main text; abilities are sampled from the standard normal distribution). For simplicity, we use the 1PL for estimation. We consider three quantities: IMV$(\bar{x}, \bar{x_j}; x^\star)$, IMV$(\bar{x}, 1\text{PL}; x^\star)$ and IMV$(\bar{x_j}, 1\text{PL}; x^\star)$. Results are shown in Figure S1. IMVs are maximized when $\mathbb{E}(b) = 0$ and decrease from there. For the IMV based on comparison to prediction from prevalence alone, IMV$(\bar{x}, \bar{x_j}) > 0.3$ at its maximum while we observe IMV$(\bar{x_j}$,1PL$)> 0.1$. These IMVs diminish to approximately 0.05 for large values of $\mathbb{E}(b)$. We view this monotonicity as reasonable behavior, given that (assuming $\mathbb{E}(\theta) = 0$) there is less uncertainty in system where $\mathbb{E}(b)$ is relatively far from zero (see discussion in Domingue et al., 2021).

Figure S1: IMV as a function of $\mathbb{E}(b)$ when the DGM and DAM are the 1PL. We simulate 500 datasets where $\mathbb{E}(b) \sim \text{Unif}(0, 1.5)$ (with $N = 1000$ respondents and 50 items) and then use LOESS to produce fitted curves.

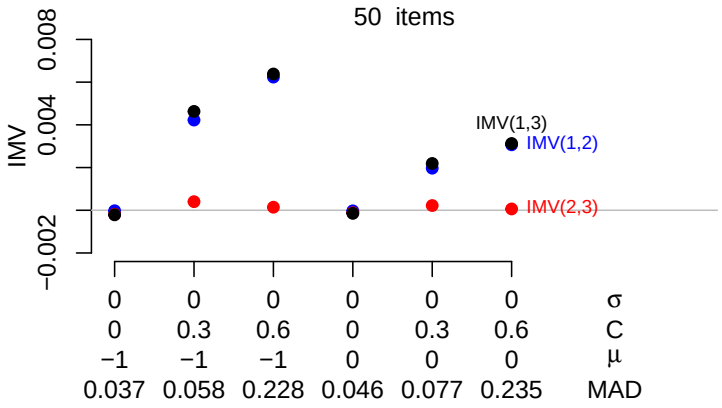


S2.2 Further analysis of the 3PL

Figure S2 further illustrates the relative lack of difference between predictions from the 2PL and 3PL even when guessing is quite pronounced (as a function of both the absolute level of guessing, C, and the overall ability, μ, of the respondents) and there are a large number of respondents ($N = 25000$). We also show the mean absolute deviation (MAD) averaged across iterations of a given configuration of generating parameters between the true and estimated guessing parameters to ensure that they are being accurately estimated. The relatively small advantage of the 3PL relative to the 2PL is shown when the red dot is slightly above the x-axis and the black dot is slightly above the blue line. Even with pronounced levels of guessing and low-ability respondents, there is relatively little benefit to be had from the 3PL; i.e., the average IMV(2PL,3PL) when $\sigma = 0$, $C = 0.6$, and $\mu = -1$ is only 8e-4. Note also that guessing parameters are poorly estimated in this case.

To further investigate whether there are differences between 2PL and 3PL item response probability estimates as a function of ability, we looked at the IMV computed separately as a function of sum score. In Figure S3, we do this for different values of C (where guessing parameters are sampled from Unif$(0, C)$) and relatively large numbers of respondents to ensure that results aren't driven by noisy estimates of guessing parameters. These results suggest that 3PL estimates are, as we might expect, somewhat more valuable for lower-ability respondees when C is relatively large, but the differences are modest.

Figure S2: IMV as a function of data-generating parameters for 50 items and $N = 25000$ respondents. σ and C are as described elsewhere. The ability of the respondents is centered at μ while item difficulties are centered at 0. We generate 100 datasets for each set of simulation conditions. The MAD describes the mean absolute difference between the true guessing parameter and the estimate across all iterations for a given set of simulations.



S2.3 The EAP versus the MLE

Using the same design as the simulation in Figure 1 of the main text, we consider here the IMV associated with usage of the EAP relative to the MLE. Results are shown in Figure S4. The EAP offers more valuable predictions in all conditions considered here. EAP estimates show the most value relative to MLEs when the 1PL is used for recovery and the DGM ("data generating model"; we similarly use DAM for "data analysis model"). However, the magnitude of the IMV is small (i.e., IMV(MLE,EAP)< 0.003) in all cases.

S2.4 The role of the prior

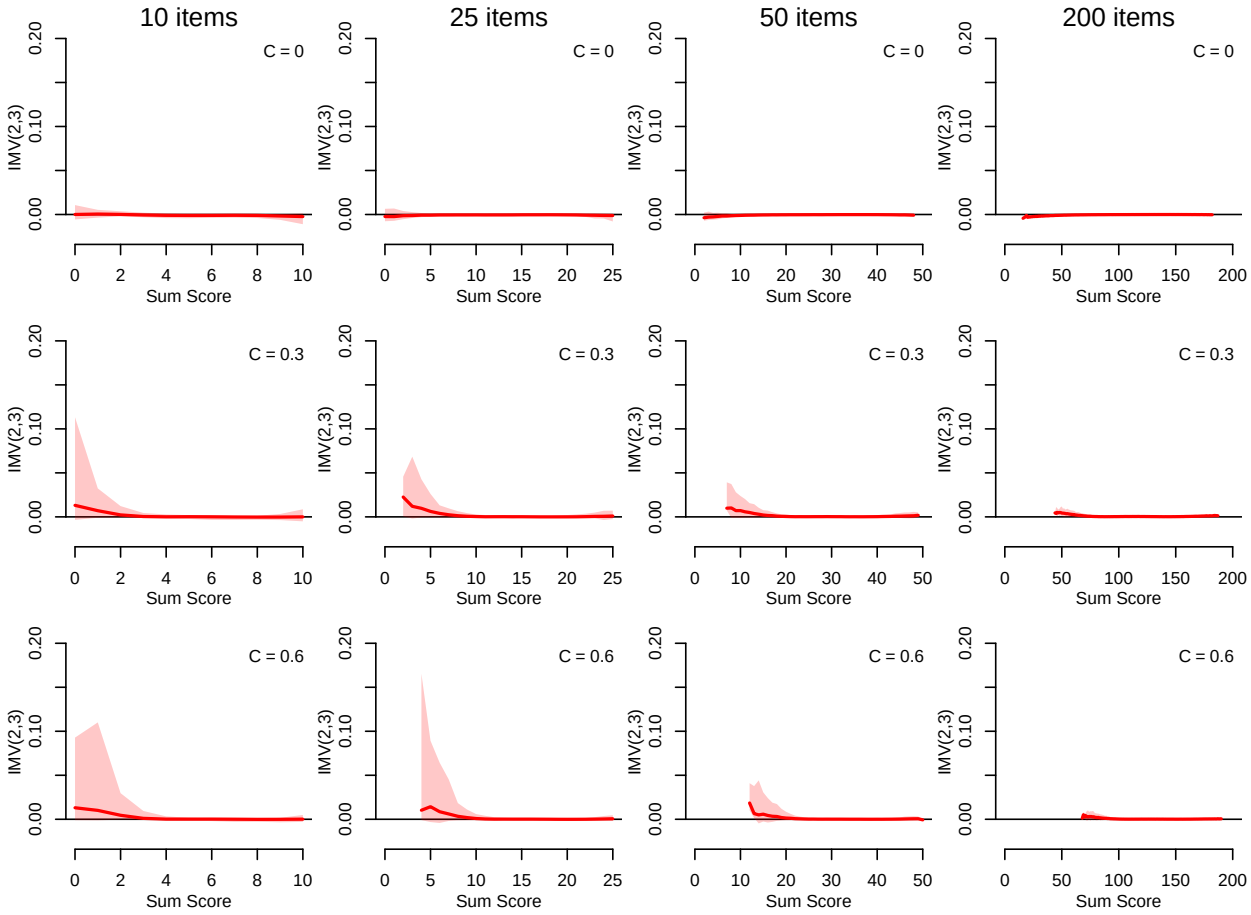
To facilitate estimation across a broad range of settings, we use priors for estimation of the discrimination and guessing parameters. Here, we describe sensitivity analyses showing the robustness of results to the choice of a prior. We first conducted a sensitivity analysis related to the prior we imposed on the discrimination parameters. We generated parameters $a_j \sim \text{LogNormal}(0, \sigma^2)$ and varied $\sigma \in \{0, 0.75, 1.5\}$. We considered priors for the a_j parameters of $\text{LogNormal}(m, s)$.[2] We let $m \in \{0, 0.2\}$ and sampled 250 values for $s \sim \text{Unif}(0.05, 1)$. We consider sample sizes of 1000 respondents and 50 items.

Results are shown in Figure S5. Results are fairly comparable as a function of m so we focus on s. When s is large ($s > 0.5$), we observe minimal differences (i.e., IMVs close to 0) between the models whether they include a prior or not. For small values of s, the IMV is positive when $\sigma = 0$ but negative when $\sigma > 0$. This is reasonable behavior; when $\sigma = 0$ there is no variation in the discrimination parameters so a hyperparameter of $s = 0$ would be appropriate while the strong assumption of a small s is costly when $\sigma > 0$. Given these results, we use $m = 0$ and $s = 1$ in analysis; such priors offer flexibility and perform reasonably in the simulation studies considered here.

We also probed the degree to which variation in the prior placed on the guessing parameter impacts the quality of subsequent estimates. We simulated data via the 3PL with guessing parameters drawn from $\text{Unif}(0, C)$ where we vary C in simulation. For estimation, we use priors of the form $\text{Beta}(2,\beta)$. Interest is in the IMV for values of β relative to $\beta = 17$ which we use in other analyses. We show the variation in the prior as a function of the choice of β in the left-hand panel of Figure S6. We emphasize here that the generating distribution of the guessing parameters does not match our choice of priors; our goal is to probe the sensitivity of resulting estimates of p_{ij} so as to understand the degree to which findings considered here may be sensitive to our specific choice of a prior.

[2]Using the parametrization here: `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Lognormal.html`.

Figure S3: IMV(2PL,3PL) computed by sum score. Results are based on 1000 simulated datasets containing 5000 people and the stated number of items (with $\sigma = 0.5$ being used to generate discrimination parameters). Red line shows median IMV for each sum score across all data while shaded region shows 0.025 and 0.975 quantiles.

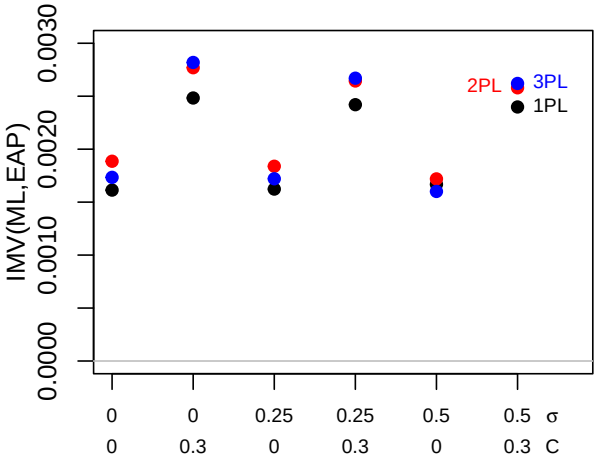


Estimation (for $N = 1000$ respondents and 50 items) are shown in the right-hand panel of Figure S6. Note that priors that place more weight on relatively large guessing parameters (e.g., $\beta \in \{1, 5\}$) perform better relative to Beta(2,17) when C is relatively large. Larger values of β tend to yield relatively little differences as a function of C with respect to their predictive performance when contrasted with Beta(2,17). However, differences are relatively small; they are, for example, roughly an order of magnitude smaller than IMVs due to switching from the ML to EAP (i.e., Figure S4). We thus conclude that the specific choice of β is unlikely to have a large impact on the key findings shown here.

S2.5 Misfit costs as a function of sample size

Here we examine various IMVs associated with different choices for the data generating process under different assumptions about the sample size. For each choice of DGM, we generated 250 datasets. In the top panels of Figure S7, we first examine IMV(1PL,2PL) and IMV(2PL,3PL) under different assumptions about the data generating model. We begin with the 1PL; in that case, there is effectively no value of the 2PL relative to the 1PL (and in fact IMV(1,2)< 0 for small number of respondents) or the 3PL relative to the 2PL irrespective of sample size. This is to be expected; when we use the 1PL to simulated data, the 2PL and 3PL should not provide additional predictive value. When the DGM is the 2PL, the IMV(1PL,2PL) is quite high for small samples and increasing as a function of sample size. When the DGM is the 3PL, the

Figure S4: IMV associated with EAP estimates of θ relative to ML-based estimates. Points show average IMVs based on different choices of σ and C. We generate 100 datasets for each set of simulation conditions.



2PL continues to generate a high IMV relative to the 1PL. The IMV(2PL,3PL) is positive but small and not increasing as a function of sample size.

The findings in the top row of Figure S7 are accompanied by those in the bottom panel which show the oracle and overfit values for the different model-derived estimates relative to the true p_{ij} values. Focusing on the oracle, the IMV associated with knowledge of truth declines as a function of sample size but only to a point (more on this in the subsequent section). Note that the oracle IMV associated with the 1PL is substantially higher when the DGM is the 2PL or 3PL than the oracle values for those models (i.e., the red line is well above the blue/black lines in second and third panels of bottom row). However, note that the overfit value for the 1PL is generally positive when the DGM is the 2PL or 3PL; this suggests a robustness to overfitting for the 1PL that we further discuss in the next section.

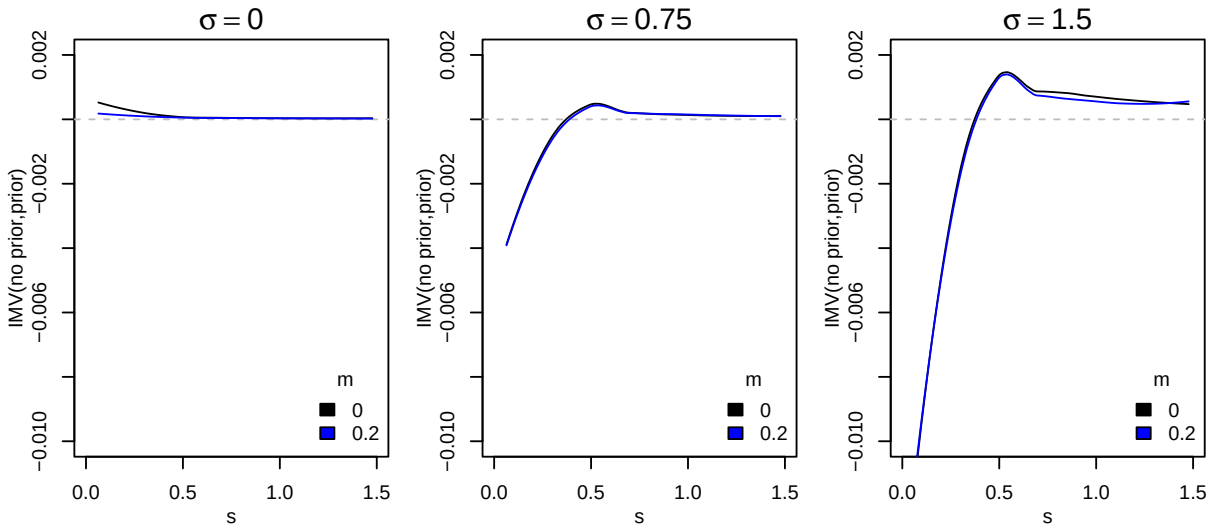
S2.6 Fit and sample size for a correctly specified model

We now use the IMV to explore the value of additional respondents and items when the appropriate model (i.e., the DGM is identical to the DAM in all cases here) is fit in terms of predictive value using the oracle. Results are shown in Figure S8 wherein we consider scenarios based on simulating data from the 1PL/2PL/3PL to 25, 50, or 200 items and then estimating the same model used to generate the data (i.e., there is no model misspecification). In the top row, we consider the oracle IMV. In all cases, we observe decreases in the oracle IMV as a function of sample size but note that, for a given sample size, the Oracle is typically smallest for the 1PL model (although differences between the 1PL and 2PL become smaller as N increases) and largest for the 3PL. Note that there is an effect of the number of items on the IMV (lines of a common color tend to be higher in panels with more items), although it is perhaps modest over the common range of items (25–50).

In Figure S8, the IMV does not go to zero for increasingly large samples but a fixed number of items. Why is this? The issue is that individual estimates of θ_i are largely unaffected by increases in the number of respondents thus leading to a floor in how accurate estimates of p_{ij} get for a given scenario. We can confirm this in the bottom panels of Figure S8 which computes the root mean squared error (RMSE) between true and estimated p_{ij} values in the test data for different sample sizes. The existence of a "floor" in the RMSE curves for a measure with a fixed number of items ultimately limits the degree to which increasing sample size increases precision of predictions.

To finish this discussion, we turn now to a second set of analyses shown in Figure S9 that emphasize the connection between the RMSE discussed above and the IMV. In this Figure, we add noise to the true response probabilities generated from the 3PL (using same distributions for difficulty, guessing, and discriminations as in Figure S8). The x-axis is the RMSE between true and noisy response probabilities; it quantifies the degree of noise. We then compute the IMV between the noisy and true estimates. Note the strong similarity

Figure S5: Role of prior for the discrimination parameter. We consider difference levels of variation in the discrimination parameters (the three panels) and examine the IMV contrasting the model with no prior to the model with the specified prior prior as a function of the two hyperparameters (m, s).



between these two quantities. In this simulated environment where truth is known, we can see that the IMV is strongly sensitive to error in the estimates of probabilities of individuals responses.

S2.7 The IMV versus alternatives

Building on Figure 3 in the main text, we consider additional simulation studies contrasting the behavior of the IMV with the RMSEA (Maydeu-Olivares, Cai, & Hernández, 2011) and the AIC (Burnham & Anderson, 2004). We continue to base simulations on the 2PL (Eqn 6 in main text with $c_j = 0$) with a_j sampled as in Section 3.4. Critically, we sample $b_j \sim \text{Normal}(\mu, 1)$ to systematically vary the prevalence—via $\mu \sim \text{Unif}(-3, 3)$—of the responses. We assume a sample size of 5000 respondents and 50 items. Results for this first simulation study are shown in Figure S10.

Consider first the RMSE (in the left panel). In absolute terms, the RMSE decreases as μ moves away from zero, given that there is less variation in the responses (i.e., for a Bernoulli random variable, the variance is a function of μ; see also SI-S2.1). The RMSE is smallest in absolute terms for the 2PL, but note the asymmetry: the 3PL is performs especially poorly when μ is large, and guessing is less salient (due to overfitting). The IMV(1PL,2PL) decreases as μ moves away from zero; this is consistent with the behavior of the RMSEs (recall also the results in Figure S9). The IMV(2PL,3PL) is negative and decreasing as μ increases. The RMSEA shows similar behavior for the 1PL but prefers the 3PL to the 2PL; this is in contrast with what we know to be the superior model based on the RMSE values and indicative of behavior from the RMSEA that may be of concern for purposes of model selection. The changes in AIC values are in the expected directions, but the lack of portability is apparent here, even absent the changes in sample size.

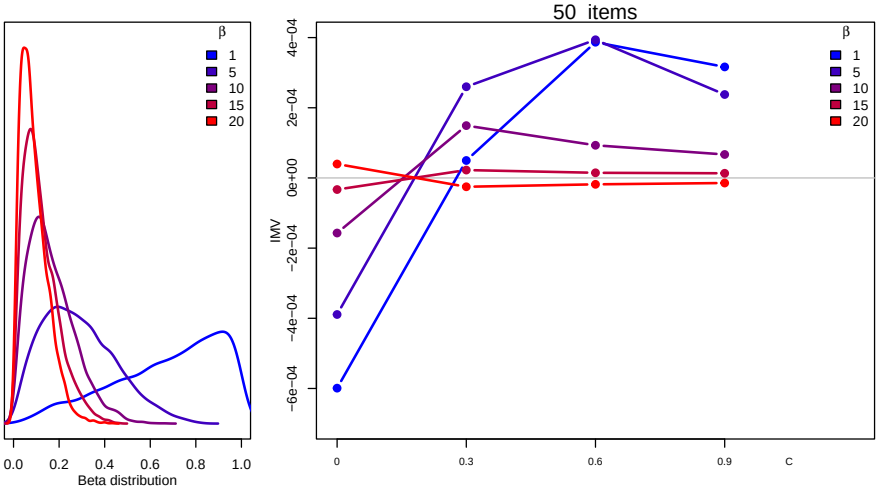
S2.8 Multidimensional models and the IMV

We conducted a simulation study related to the IMV's performance with multidimensional IRT so as to offer context for the empirical analyses with multidimensional IRT models. We use a standard compensatory mirt model; i.e., we let

$$p_{ij} = \frac{1}{1 + \exp(\boldsymbol{a}_j \cdot \boldsymbol{\theta}_i + b_j)} \tag{S1}$$

where $\boldsymbol{a}$ and $\boldsymbol{\theta}$ are vectors of dimension K (here we set $K = 2$). We sample $\boldsymbol{\theta}_i$ from a multivariate normal distribution with zero-mean, unit variances, and a covariance of ρ; we sample b_j from a standard normal distribution. We sample elemnts of $\boldsymbol{a}_j$ from $\text{LogNormal}(0, 1^2)$ and then use a parameter $\tau \in [0, 1]$ to moderate

Figure S6: Role of prior for 3PL guessing parameter. Left: Illustration of Beta(2,β) for different choices of β. Right: Average IMV for Beta(2,β) relative to Beta(2,17) as a function of C. We simulated 25 datasets for each configuration of simulation conditions.



the degree to which dimensionality is within- or between-item. If there are N_j items, we sample a proportion (τN_j) of items and, within that portion, randomly set one element of $\boldsymbol{a}_j$ to zero; when $\tau = 1$, dimensionality is fully between-item whereas when $\tau = 0$ it is fully within-item. We sample 100 values for ρ and choose $\tau \in \{0, 0.5, 1\}$ (and fix the number of items $N_j = 50$ and the number of respondents $N = 5000$); results are smoothed across values of ρ.

We estimate the same compensatory model using `mirt` (Chalmers, 2012). We allow for correlations between the latent factors and also impose priors on both loadings using the same prior as in the 2PL case. Results based on variation in τ and ρ are shown in Figure S11. In the left panel we show the IMV(2PL,2F-2PL) as a function of ρ; each line is based on results for for a separate value of τ. As ρ increases, the model becomes effectively unidimensional and there is (as expected) generally little value in the 2F-2PL approach relative to the 2PL. When ρ is small, the relative value of the 2F-2PL approach depends upon τ. The predictive value of the 2F-2PL is greater when τ is larger (i.e., when multidimensionality is between-item); this is reasonable given that the 2PL will tend to misfit items more substantially where one loading is set to zero. In the right panel, we complement the IMV results with results based on the RMSE given that the true response probabilities are known (i.e., we compute the RMSE across p_{ij} from Eqn S1 versus $\widehat{p_{ij}}$ from one of the two models). The fact that the estimates of p_{ij} become worse for the 2PL as $\rho \to 0$ is readily apparent.

Figure S7: The IMV associated with different item response models (top) and the oracle/overfit values (bottom) for the 1PL/2PL/3PL for different choices of data generating model (DGM). We simulated data for 250 choices of $n \sim \text{Unif}(2, 4)$ where the number of respondents was 10^n with $C = 0.3$ for the 3PL and $\sigma = 0.5$ for the 2PL and 3PL (we focus on 50 items throughout). We focus on LOESS regressions of the resulting IMVs as a function of $\log_{10}$ of the sample size.

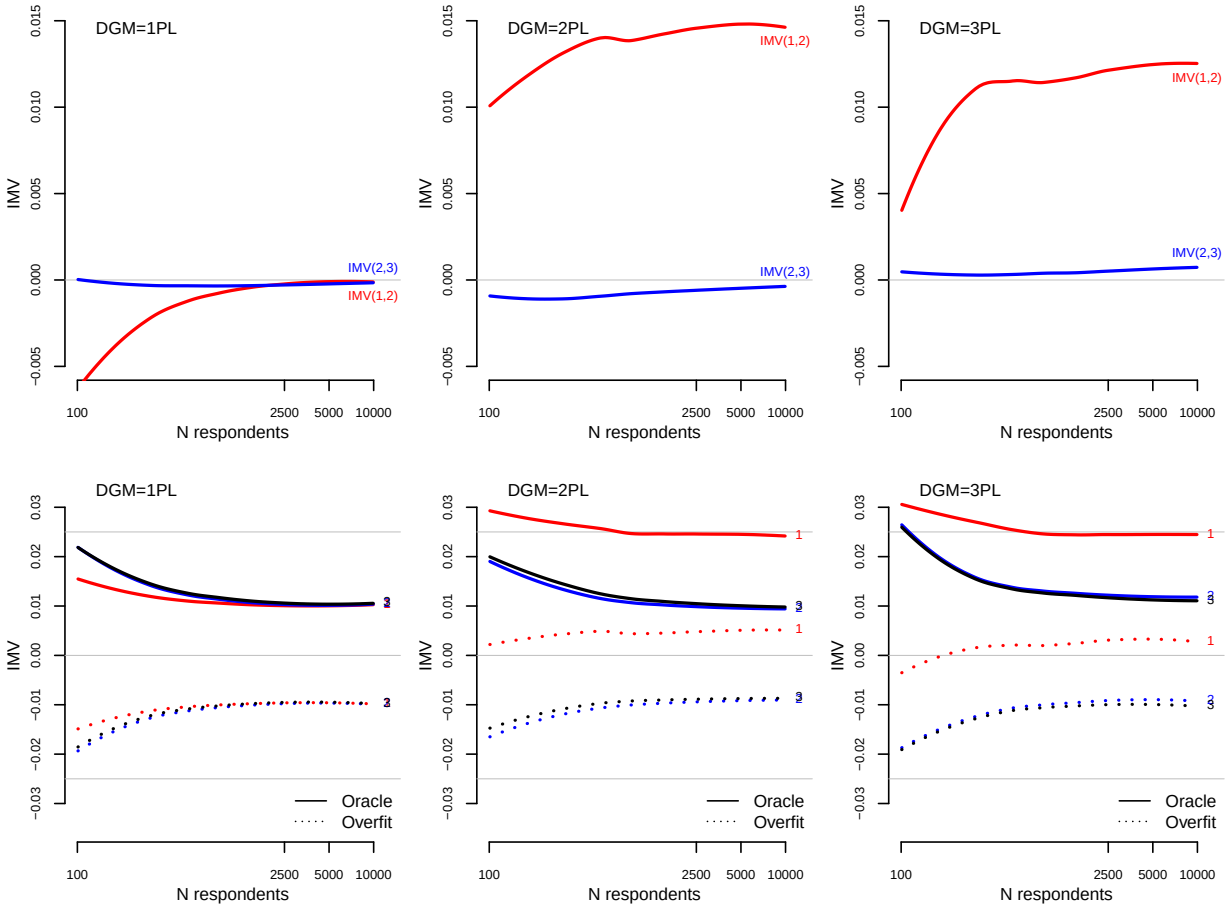


Figure S8: The value of sample size. For each choice of DGM, we simulated data for 250 choices of $n \sim \text{Unif}(2, \log_{10} 5000)$ where the number of respondents was 10^n. For the 3PL we chose $C = 0.3$ and $\sigma = 0.5$ for the 2PL and 3PL. Resulting curves based on LOESS regression. Top: The oracle IMV associated with IRT-based estimates. Bottom: Oracle and Overfit IMV estimates.

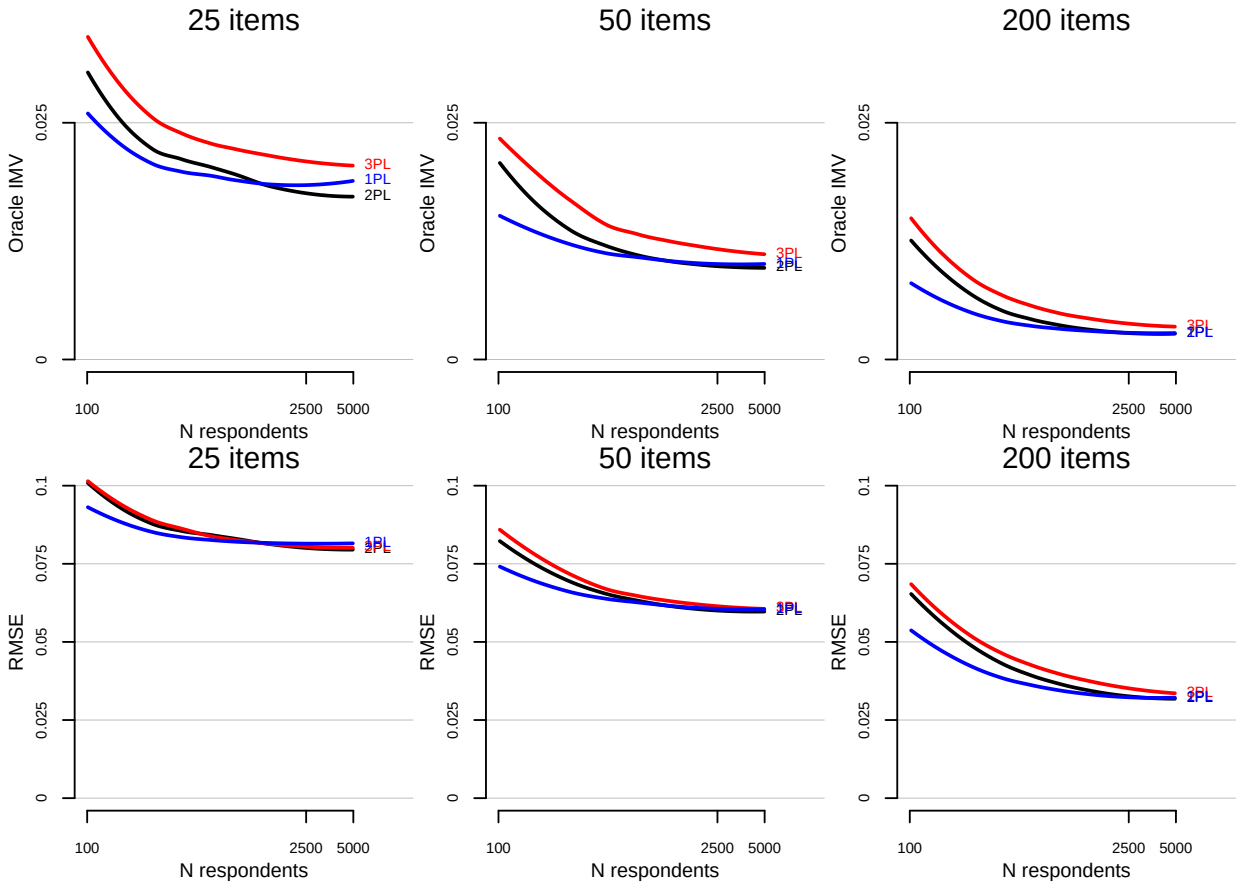


Figure S9: IMV for true 3PL p_{ij} (with $C = 0.3$ and $\sigma = 0.5$) values versus those values observed with noise for $N = 2000$ respondents. For a given response with true probability p_{ij}, we consider noisy probabilties of $p_{ij} + \delta_{ij}$ with $\delta_{ij} \sim \text{Unif}(-\Delta, \Delta)$ (and the caveat that if $p_{ij} + \delta_{ij} > 1$ we censor at $1 - \epsilon$ and if $p_{ij} + \delta_{ij} < 0$ we censor at ϵ for $\epsilon = 0.001$). We consider 25 iterations based on different choices of $\Delta \sim \text{Unif}(0,.2)$.

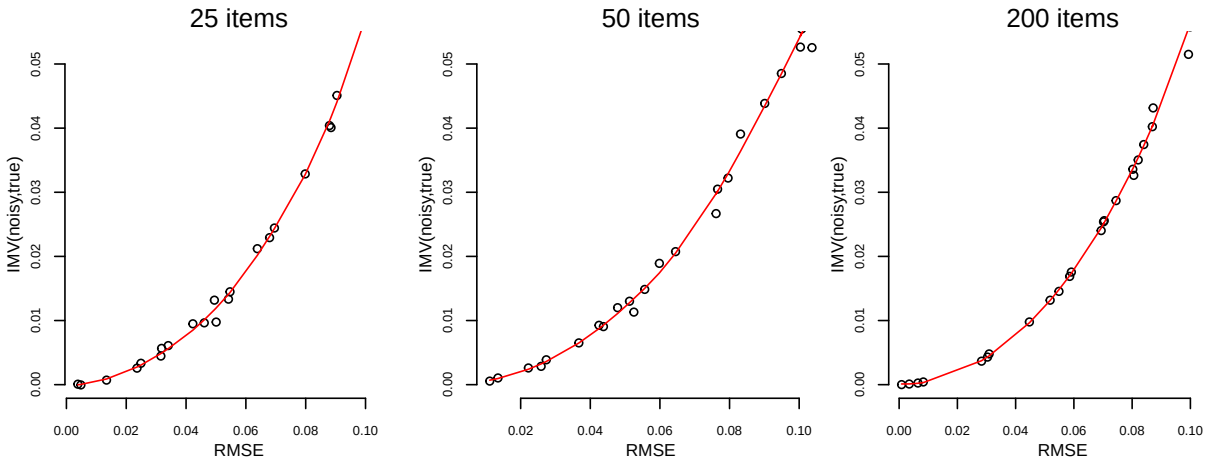


Figure S10: Simulations comparing a variety of metrics (in columns; along with RMSE as compared to the true/known probabilities used to generate item responses) for 1/2/3PL estimates (shows as different colors). Data are generated via the 2PL. Solid lines indicate comparisons; in the first row, the dashed lines indicate raw RMSE for the 3 approaches (1PL, black; 2PL, blue; 3PL, red). Results are based on LOESS smoothing for 250 choices of μ.

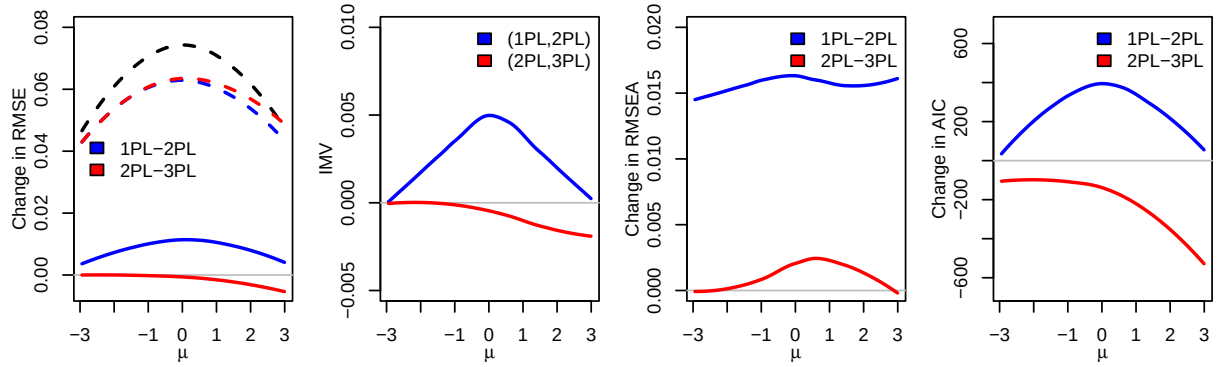
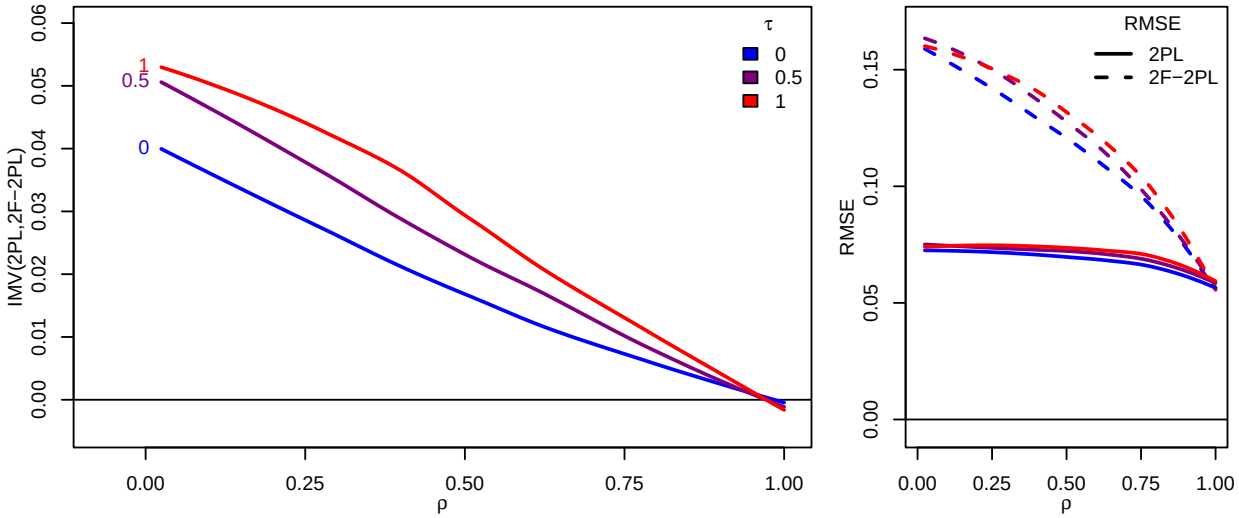


Figure S11: Results from multidimensional simulations. The left panel focuses on IMV(2PL,2F-2PL) as a function of ρ and τ. The right panels show the RMSE between the true and estimated probabilities separately for the 2PL and 2F-2PL models. Results are based on LOESS smoothing for 200 choices of ρ.



S3 Empirical Data

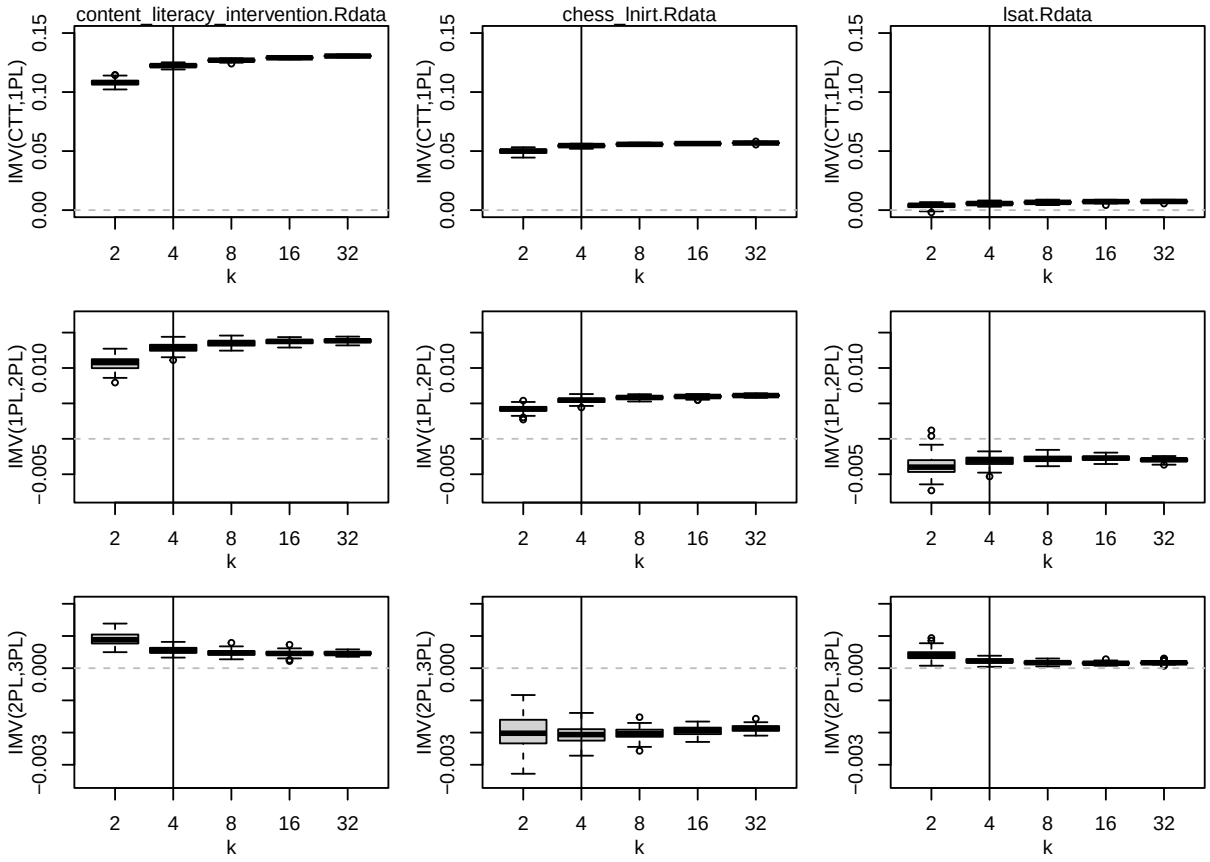
S3.1 Description of Data

We use empirical data from 89 datasets. Much of the data comes from the publicly available resources of the Item Response Warehouse (IRW, Domingue & Kanopka, 2023) but some datasets cannot be publicly reshared due to licensing restrictions. A file containing information about all of the data used here is available as a separate supplemental document; this file also contains dataset-specific IMV results.

S3.2 Sensitivity to the number of folds

We conducted a small study to probe the sensitivity of findings to the number of folds k used in cross-validation with three empirical datasets. Results are in Figure S12 where we look at the sensitivity of IMVs to choices of $k \in \{2, 4, 8, 16, 32\}$. For each choice of k, we computed the average IMV across k folds; we did this 100 times and show boxplots for these 100 values over each choice of k. The IMV values show minimal variation to different choices of k.

Figure S12: Sensitivity of IMV values to number of folds k for three datasets (we use $k = 4$ for empirical analyses in main text).



References

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, *33*(2), 261–304.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the r environment. *Journal of statistical Software*, *48*(1), 1–29.

Domingue, B., & Kanopka, K. (2023). The item response warehouse (irw).

Domingue, B., Rahal, C., Faul, J., Freese, J., Kanopka, K., Rigos, A., ... Tripathi, A. (2021). Inter-model vigorish (imv): A novel approach for quantifying predictive accuracy when outcomes are binary. Retrieved from `https://osf.io/gu3ap/`

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2023). Modeling item-level heterogeneous treatment effects with the explanatory item response model: Leveraging large-scale online assessments to pinpoint the impact of educational interventions. *Journal of Educational and Behavioral Statistics*, 10769986231171710.

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(3), 333–356.