The InterModel Vigorish as a lens for understanding (and quantifying) the value of item response models for dichotomously coded items

Benjamin W. Domingue[1,†], Klint Kanopka[1], Radhika Kapoor[1], Steffi Pohl[2], R. Philip Chalmers[3], Charles Rahal[4], and Mijke Rhemtulla[5]

[1]Graduate School of Education, Stanford University
[2]Freie Universität Berlin
[3]York University
[4]University of Oxford
[5]University of California, Davis
[†]Correspondence about the paper should be sent to ben.domingue@gmail.com.

Accepted for Publication at Pscyhometrika. Please cite this as:

Domingue, B. W., Kanopka, K., Kapoor, R., Pohl, S., Chalmers, R. P., Rahal, C., & Rhemtulla, M. (2024). 'The InterModel Vigorish as a lens for understanding (and quantifying) the value of item response models for dichotomously coded items'. *Psychometrika*, 1-21.

Acknowledgements
This work was supported in part by the Jacobs Foundation (BD), and the Leverhulme Trust (Grant RC-2018-003) for the Leverhulme Centre for Demographic Science (C.R.). Some of the computing for this project was performed on the Sherlock cluster at Stanford University. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results. The authors would like to acknowledge helpful comments from Leah Feuerstahler, Scott Monroe, David Torres Irribarra, Roy Levy, and the members of the Measurement Lab at the Harvard Graduate School of Education

.

Competing Interests
The authors have no competing interests to declare that are relevant to the content of this article.

Data availability
This paper relies on a mixture of public and private data. Links to public data are shared in the SI. Code is also available, see SI.

Abstract

The deployment of statistical models—such as those used in item response theory (IRT)—necessitates the use of indices that are informative about the degree to which a given model is appropriate for a specific data context. We introduce the InterModel Vigorish (IMV) as an index that can be used to quantify accuracy for models of dichotomous item responses based on the improvement across two sets of predictions (i.e., predictions from two item response models or predictions from a single such model relative to prediction based on the mean). This index has a range of desirable features: it can be used for the comparison of non-nested models and its values are highly portable and generalizable. We use this fact to compare predictive performance across a variety of simulated data contexts and also demonstrate qualitative differences in behavior between the IMV and other common indices (e.g., the AIC and RMSEA). We also illustrate the utility of the IMV in empirical applications with data from 89 dichotomous item response datasets. These empirical applications help illustrate how the IMV can be used in practice and substantiate our claims regarding various aspects of model performance. These findings indicate that the IMV may be a useful indicator in psychometrics, especially as it allows for easy comparison of predictions across a variety of contexts.

1 Introduction

The utilization of statistical models for item responses gathered in psychological assessments necessitates tools that describe their relative performance in a given scenario. A wide variety of indices for quantifying the quality of models for item responses exists (for a recent review, see Chapters 17–20 in Van der Linden, 2017a). However, while there are many popular approaches, they tend to have limitations. In particular, many indices do not have values that readily generalize across samples. In some cases, this can be due to a dependency on sample size. In other cases, it can be due to a poorly understood sensitivity to item difficulty. Further, such indices may not be readily applicable in scenarios wherein interest is in out-of-sample prediction; an important shortcoming given the increased relevance of prediction in many settings (Rahal, Verhagen, & Kirk, 2022; Watts et al., 2018; Hofman et al., 2021).

When using predictive models for dichotomous item responses, a metric that is portable—i.e., its values can be consistently and meaningfully interpreted to evaluate the predictive value of different models for item responses—would be valuable as it would allow for the comparison of a variety of modeling choices in various data contexts. We also distinguish between portability and the metric's sensitivity to changes in predictive accuracy. The metric needs to be sensitive to factors that change the accuracy of predictions. For simplicity, we articulate this distinction via a consideration of sample size in the linear regression context. Consider a simple linear regression model in which we are predicting some outcome y via $\hat{y} = \mathbb{E}(y|x)$. In such a scenario, we anticipate better predictions as sample size increases (see discussion below Eqn 6.35 on p.210 of Wooldridge, 2013). In this sense, we would anticipate the difference $|\,\mathbb{E}(y|x) - y|$ to get smaller as the sample size increases. However, the differences $|\,\mathbb{E}(y|x) - y|$ are not portable given that they depend on the scale of y. In the linear regression context, it would perhaps be sufficient to rescale y. Binary outcomes are more complex given that both the mean and variance depend upon the single parameter of a Bernoulli random variable. For binary outcomes such as dichotomous responses, the InterModel Vigorish (IMV; Domingue et al., 2021) is designed to resolve this problem of portability given its construction and we illustrate its appropriate sensitivity to factors that improve the prediction of novel responses in various simulation studies below.

Here, we build on initial work (i.e., the results in Domingue et al., 2021) to showcase how the IMV can be used in psychometric settings. In this paper, we conduct a series of simulation studies showing how the IMV can be used to understand the differences in predictions derived from IRT models for dichotomous items under a number of conditions. In addition, we leverage the portability of the IMV to make direct comparisons that allow us to describe the degree to which specific modeling choices impact prediction in controlled settings. We study the implications of a wide range of choices—the differences of distributions of key model parameters, of sample size, Bayesian priors, and estimation algorithms—on prediction quality in a metric that is both portable across these settings and whose values can be readily extended to work with empirical data. To illustrate this last point, we

conduct empirical work involving a large volume of data (89 datasets from the Item Response Warehouse (IRW); Domingue & Kanopka, 2023). These empirical analyses demonstrate the utility of the IMV in practice and assess the degree to which the modeling innovations considered in the simulations lead to predictive gains anticipated in idealized settings.

We also study the behavior of the IMV vis-à-vis the behavior of other alternative indices (e.g., information criteria; Burnham & Anderson, 2004). To be clear about our expectations: we anticipate that all indices are likely to provide similar information if the objective is to simply determine whether one model is a 'better' fit to data than another. Our interest here is in what these indices tell us about the differences between models in relative rather than absolute ways (i.e., how much better is one model than another?). The work in this paper builds on initial simulation studies conducted with the IMV (Domingue et al., 2021). Evidence from that paper suggests that the IMV is quite sensitive to estimation error in a way that other indices (e.g., AUC; Hanley & McNeil, 1982) are not. These differences suggest that when there is interest in more than simply ranking the performance of models, the IMV provides novel information (that is closely related to errors in prediction) which may be useful in understanding the quantitative differences in their predictions. In this paper, we complement those findings with additional results focusing on indices widely-used with latent variable models. These new results focus closely on the issue of sample size and its association with prediction quality.

This paper is organized as follows. We first discuss other indices before introducing the IMV in the context of item response models for dichotomously scored item responses. We then evaluate its performance in a variety of simulations both in isolation and in comparison to other indices. The metric's applicability to empirical data is then illustrated in a wide variety of datasets. We close with a discussion of the IMV's potential use in psychometrics.

1.1 Conventional Fit Indices for IRT models

There is a substantial literature on assessing the degree to which a given IRT model aptly characterizes a given set of item responses. Rather than a complete review, we focus on key points related to the index we develop here. Many approaches involve computation of quantities for the purpose of assessing the data-model match. We will generically call these quantities "indices". There are several types of indices to consider as alternatives (Swaminathan, Hambleton, & Rogers, 2006), some of which are meant to interrogate specific assumptions of the relevant IRT model. These include, for example, the infit and outfit statistics associated with the one parameter logistic (1PL) model (Wu & Adams, 2013) and analyses of dimensionality (Stout, 1987); we do not further discuss such indices here. There is also research on item- (Köhler, Robitzsch, & Hartig, 2020) and person-level (see Chapter 6 in Van der Linden, 2017b) fit indices; we do not focus on such indices but return to this issue in Section 5. Rather, we focus on a range of approaches meant to describe and compare the overall "fit"—by

which we mean, roughly, "how close are predictions to observations?"—of a given model to a dataset. We include discussion of indices that have distinctive features but that are similar in the sense that they are potential tools for the job of selecting or differentiating amongst various models.

One approach includes classical inferential tests of differences between models based on the likelihood ratio test. This is a widely-used approach but hinges on the availability of large samples (Mavridis, Moustaki, & Knott, 2007). In finite samples, there is frequent interest in likelihood-based approaches that correct for overfitting by favoring parsimony (Kang & Cohen, 2007). Such indices—e.g., Akaike's information criterion (AIC; Akaike, 1973), Schwarz's Bayesian information criterion (BIC; Schwarz, 1978), and the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002)—compare the fit of nested and non-nested models (including null models). One challenge with utilization of these indices is that their values are dependent on sample size; while approaches exist that reduce this dependence (e.g., Wagenmakers & Farrell, 2004), common usage hinges on values that are non-portable due to this sample size dependence.

A different approach involves examination of the contingency tables. For a test with n dichotomous items, there are 2^n possible response patterns. A complete comparison of the observed versus expected responses of each pattern would be computationally burdensome for even moderate n, but approaches emphasizing lower-dimensional summaries are useful (Maydeu-Olivares & Joe, 2005). Resulting statistics such as the M_2 can also be converted to root mean square error of approximation (RMSEA) type indices that also emphasize model parsimony (Maydeu-Olivares, 2013). The RMSEA is not sample size dependent and is useful as a measure for overall fit of a model to the data but is challenging to use for purposes of comparison, as differences in RMSEA are sensitive to the size of the initial model (in degrees of freedom) which may lead to an inability to detect misfit in differences between large models (Savalei, Brace, & Fouladi, 2021). In this paper, we compare the IMV to the AIC and RMSEA. These quantities differ in key ways, but they are all regularly used for the purpose of evaluating the performance of the kinds of models we consider here and they are thus useful for the purpose of evaluating the IMV. However, the IMV is also informed by recent shifts in thinking on the importance of prediction. We discuss these shifts below.

1.2 From explanation to prediction

Traditionally, fit indices and model parameters have been computed based on the same data. Historical limitations on computation frequently mandated such approaches and, of course, indices were often designed with such limitations in mind (e.g., the AIC penalty for overfitting based on the number of estimated parameters, Stone, 1977). More recent thinking, however, emphasizes the advantages of evaluating fit based on out-of-sample data in psychology (Yarkoni & Westfall, 2017), across the social sciences more broadly (Verhagen, 2022; Wolfram, Tropf, & Rahal, 2022), and into the computational sciences more generally (Savcisens et al., 2023). This rapidly occurring

(Rahal et al., 2022) change in perspective is tied to criticisms that social science research has historically been too narrowly focused on finding causal mechanisms based on an in-sample analysis of association-based models applied to observational data (Shmueli, 2010). As one example from psychological measurement of the gains such approaches may offer as compared to conventional in-sample studies, out-of-sample approaches may allow for improved identification of dimensionality in factor analysis settings (Haslbeck & van Bork, 2024).

We agree with arguments that such a consideration of prediction is essential for improving our theoretical understanding even when there is no inherent interest in prediction itself (Watts et al., 2018; Hofman et al., 2021). The move to prediction—not to be conflated with the use of highly bespoke models for use in 'forecasting' exercises (Watts, 2014)—allows us to provide improved insight into model fit, to construct bench-marking tools across modeling domains, and to generate insight into the behavior of complicated models. Full enjoyment of these benefits may require the use of novel indices for understanding the performance of predictive models above and beyond conventional indices that are perhaps not well-suited to this purpose.

Thus, prediction of out-of-sample data has begun to emphasize indices specifically designed for such purposes. An early example of this kind of analysis in psychometric settings emphasized a version of the out-of-sample log-likelihood as performing better than many alternatives (Kang & Cohen, 2007).[1] A more recent paper (Stenhaug & Domingue, 2022) introduced one insight that is key for our purposes. Frequently, analysis of out-of-sample data for purposes of model selection has focused on which approaches allow one to identify the data generating model. In contrast, they (Stenhaug & Domingue, 2022) argue that we should instead be asking which models are maximally predictive of out-of-sample data. The more predictive model should be favored, irrespective of whether this model is also the data generating model. Although we would clearly expect them to be in some cases, there are other cases where the data generating model may fare poorly for prediction (e.g., a highly complex model may generate poor predictions relative to a simpler alternative if there is insufficient data for precise estimation of the many parameters of the complex model). In the sense that it is designed to gauge the quality of prediction in out-of-sample tests, the IMV is a "predictive index". This nomenclature is introduced so as to distinguish the IMV and its computation from more conventional indices such as those discussed above; this kind of predictive index is meant to help in quantifying modeling progress (Watts et al., 2018) with the ultimate goal of providing a 'solution-oriented' approach to social science (Watts, 2017).

[1] This idea dates back more broadly to the Generalized Cross-Validation (GCV) method (Craven & Wahba, 1978) used to estimate the correct degree of smoothing noisy data with spline functions.

2 The InterModel Vigorish for IRT models of dichotomous outcomes

2.1 The InterModel Vigorish

The IMV was first introduced in the context of generic dichotomous outcomes; we briefly describe its computation (for additional details, see Domingue et al., 2021) before moving to a discussion of its use in IRT settings. Consider the likelihood assigned by some model to each predicted outcome $p_i \in (0, 1)$ for some Bernoulli random variable $y_i \in \{0, 1\}$ (with $i \in \{1, ..., n\}$):

$$L_i = p_i^{y_i}(1 - p_i)^{1-y_i}. \tag{1}$$

We can summarize these via the geometric mean of the likelihoods

$$A = \left(\prod_{i=1}^{n} L_i \right)^{\frac{1}{n}}. \tag{2}$$

The IMV is based on a sequence of bets involving coins; we now note how these coins are identified before describing their usage. We identify a coin of weight w via a calcluation involving A; specifically, for a predictive system that leads to A, we find $w \in [0.5, 1]$ such that

$$w \log(w) + (1 - w) \log(1 - w) = \log A. \tag{3}$$

The coin with weight w has uncertainty equivalent to that of the full set of predictions p_i of some outcome; a weight close to $w = 0.5$ indicates a predictive system with high levels of uncertainty whereas a coin with weight close to $w = 1$ indicates predictions that are much more deterministic.

Suppose we now have predictions of the outcomes y from two models; we will denote these predictions as p_0 and p_1 (omitting the i subscript). Using Eqn 3 we identify coins w_0 and w_1. A fair bet (in the sense that neither side expects profit) is established based on w_0; one player bets \$1 on the positive outcome while a second bets $\frac{1}{O}$ on the negative outcome ($O = \frac{w_0}{1-w_0}$). Unbeknownst to the player betting on the negative outcome, the coin of weight w_0 is replaced with a coin of weight w_1. If $w_1 > w_0$ (i.e., predictions p_1 are better than those of p_0), the player betting on the positive outcome stands to gain. The IMV is this gain; it is based on the expected profit for the player betting on the positive outcome if w_0 is replaced with w_1. In that case, the person betting on the positive outcome now has additional information and expects to win

$$\text{IMV} \equiv \frac{w_1 - w_0}{w_0}. \tag{4}$$

This quantity is the expected profit associated with the side information contained in the p_1 prediction that is

only available to one party in a bet while the other party only has information contained in the p_0 prediction.

We pause to emphasize one crucial fact. Given that the calculations in Eqn 3 are based on the unadjusted likelihood, the IMV will be biased in favor of more complex models when evaluated in-sample. We thus rely on computation of the IMV in data not used for model estimation throughout.

The IMV has several favorable properties. First, it is a generalizeable metric that is comparable across different data and models and, thus, can be used to generalize results across applications. Generalizability is ensured given that the IMV is always conditioned on the fair bet involving w_0. Second, given that IMV always requires predictions from two approaches[2], it naturally indexes change between the approaches. However, there is a natural null model (i.e., the outcome's prevalence) that makes the IMV appropriate for evaluating the performance of a single predictive approach. Third, the IMV requires only predictions from two models about data; there are few additional restrictions. It can thus be used to make a variety of comparisons: We make comparisons across different item response models applied across different datasets, but it can also be used to make comparisons between different structural conditions (e.g., sample size), estimation strategies, and even between truth and estimates. We attempt to illustrate this flexibility—e.g., with use of the "Oracle" analysis introduced below meant to capture the last point—throughout the remainder of the paper. Fourth, values of the IMV can be interpreted straightforwardly as real numbers given their derivation. For example, if an IMV value from one predictive exercise is 10 times the value of the IMV from another, we can say that the information in the first exercise provides an order of magnitude more predictive value than that in the second exercise. We now discuss the extension of the IMV approach to an IRT framework.

2.2 The IMV for IRT models with dichotomous outcomes

When considering dichotomous item responses, application of the IMV is a fairly straightforward extension of the approach described above. Suppose we want to evaluate the fit of an IRT model to the dichotomously coded item response $x_{ij} \in \{0, 1\}$ of person $i \in \{1, \ldots, N\}$ to item $j \in \{1, \ldots, J\}$. We consider item response models that describe the probability of a correct response for person i to item j,

$$\Pr(x_{ij} = 1) \equiv p_{ij}. \tag{5}$$

If, for example, we are considering the 3PL (Lord & Novick, 1968) then the predicted probability of a correct response is modeled as a function of person ability θ_i, item difficulty b_j, item discrimination a_j, and guessing parameter c_j:

$$p_{ij} = c_j + \frac{1 - c_j}{1 + \exp(-a_j(\theta_i - b_j))}. \tag{6}$$

[2]In the sense that the IMV can be used as a relative metric, is is similar to, for example, the Tucker-Lewis index which has been used in structural equation modeling (Maydeu-Olivares & Garcia-Forero, 2010; Han, Zhang, Jiang, & Shi, 2022; Cai, Chung, & Lee, 2021).

We now introduce subscripts to denote probabilities from different approaches (and suppress i and j subscripts for readability). The vector of response-level probabilities, p_1, for the model of interest is constructed relative to some baseline model whose probabilities we denote as p_0. The value p_0 could be the predicted probability of a correct response from an alternative item response model—e.g., the 1PL model ($\forall j, a_j = 1, c_j = 0$) or the 2PL ($\forall j, c_j = 0$)—if p_1 is based on the 3PL. As a baseline, we can even consider simpler alternatives such as the overall mean ($\bar{x}$) or item-specific predictions that ignore information about the respondent ($\bar{x_j} \equiv \frac{1}{N}\sum_{i=1}^{N} x_{ij}$; i.e., the item p-value from classical test theory, Crocker & Algina, 1986).[3]

Alongside predictions p_0 and p_1, the computation of the IMV requires a specific set of outcomes but is flexible in that any set of outcomes/predictions are sufficient. If outcomes are denoted as x, then we denote the IMV metric as IMV$(p_0, p_1; x)$. Here we present out-of-sample predictions averaged across all items but emphasize this flexibility upfront. In computation of the IMV, we will use data x that are not included in the process of estimating p_0 or p_1 (i.e., x is out-of-sample or test data).[4] We occasionally denote such data as $x^\star$ when we want to emphasize that it is out-of-sample but retain the simpler x notation here to emphasize that the basic idea does not require out-of-sample data.

The IMV metric has various advantages. First, it can be used as an index in relative isolation (comparing model-based predictions to, for example, overall difficulty or item-specific difficulty) or as a comparison between two more sophisticated models. Second, when used in this second, relative sense, it offers great flexibility in the choice of comparison model (which does not need to be nested, and may also use different estimators). Third, being standardized as the expected profit of a bet, it is comparable across different models and data sets. In the following section we discuss the performance of the IMV in a range of simulation studies but also offer simple illustrations of how to compute the IMV using both simulated and empirical data (see SI-S1).

3 The IMV in simulation studies

In this section we illustrate the performance of the IMV in the context of dichotomous item responses using simulation studies. We specifically (1) evaluate how the IMV behaves under model misspecification, (2) demonstrate how it can be used to study overfitting and predictive accuracy, and (3) assess sensitivity to sample size. We also (4) contrast the behavior of the IMV to that of alternative metrics. Finally, we (5) contrast the range of IMVs computed in the simulation studies.

Note the following details regarding the simulation studies:

[3] Note that the $\bar{x_j}$ value is also termed the 0PL in some settings. However, there is ambiguity in how this term is used with some usage indicating predictions invariant across persons (as used here; see the '0PL-item' model in Reddy, Labutov, Banerjee, & Joachims, 2016) whereas in other cases it indicates predictions invariant across items (Haberman, Sinharay, & Lee, 2011; Wainer, 2016).

[4] In simulation settings, we will generate new data for testing purposes from the known data generating model. In empirical settings, we use K-fold cross-validation (James, Witten, Hastie, Tibshirani, et al., 2013).

- Unless otherwise noted, we simulate data x via Eqn 6 with $\log a_j \sim \text{Normal}(0, \sigma^2)$, $b_j \sim \text{Normal}(0, 1)$, and $c_j \sim \text{Unif}(0, C)$ for $N = 1000$ respondents and varying numbers of items $N_j \in \{10, 25, 50, 200\}$. We sample abilities $\theta_i \sim \text{Normal}(0, 1)$. For each condition, we generate 100 data sets.

- Item response models are estimated with `mirt` (Chalmers, 2012). Item parameters are estimated via the expectation-maximization (EM) algorithm; person abilities are estimated via expected a posteriori (EAP). So as to ensure convergence in the case of small samples, where applicable we estimate item parameters using a lognormal prior with parameters (0,1) for discriminations and a beta prior for guessing with parameters (2,17).[5]

- We compute the IMV using out-of-sample responses. In simulations, we produce a test dataset by sampling a new set of item responses, $x^\star$, based on the true p_{ij} values. That is, we generate new responses from the same probability distribution used to generate the training data from which model estimates are derived; note that we thus compute the IMV based on the same amount of data used for estimation. Predictions associated with these out-of-sample responses are based on the estimated item- and person-level parameters for a given models used for estimation.

Note that we use a similar approach to estimation in our empirical work discussed in Section 4.

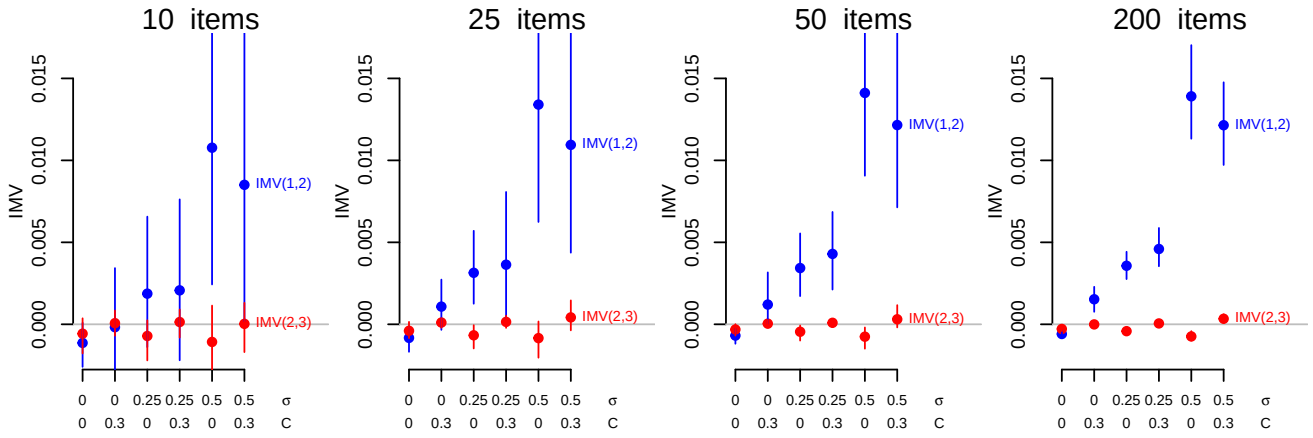
3.1 Prediction and model misspecification

We begin by investigating the performance of the IMV in analyzing estimates from models that do not necessarily have the same form as the generative model. To that end, we simulated data x using Eqn 6 (based on manipulation of $C \in \{0, .3\}$ and $\sigma \in \{0, .25, .5\}$). All conditions were fully crossed. Based on the true response-level probability of a correct response, we generate an equivalently sized holdout sample $x^\star$ of responses. We then fit the 1PL, 2PL, and 3PL models to the data in x, obtaining estimates of $\widehat{p_{ij}}$ from each and then compute IMVs for different combinations of predictions of $x^\star$. In particular, we consider IMV(1PL,2PL;$x^\star$) and IMV(2PL,3PL;$x^\star$) as a function of the σ and C values. We hypothesize, for example, an increase in IMV(1PL,2PL;$x^\star$) as σ increases. We are able to make explicit statements about this hypothesis by quantifying the value associated with fitting more complex models.

Results are presented in Figure 1 for varying numbers of items. We begin by first considering the blue points which show IMV(1PL,2PL). As expected, the 2PL provides increasing value as σ increases relative to the 1PL. For $\sigma = 0.5$, the IMV is between 0.01 and 0.015. We provide IRT-specific context for these values in subsequent sections; but, as an initial benchmark, evidence from other settings suggest that the move from the 1PL to 2PL is as valuable—in IMV terms—as, for example, information about age in prediction of chronic disease amongst older people (Domingue et al., 2021). We view these values as evidence that the IMV is clearly able to detect

[5] We considered the sensitivity to choice of prior in simulation studies, see SI-S2.4.

Figure 1: The cost of misfit: IMV values for the 3PL relative to the 2PL (red) and the 2PL relative to the 1PL (blue). Points are averages across 100 datasets for each configuration of parameters (1000 respondents in all cases); line segments represent span of 0.025 to 0.975 quantiles over the 100 datasets. The "(N,M)" in parentheses are shorthand for the N and M parameter logistic IRT models.



scenarios wherein an overly restrictive model is being used as compared to a model that generates more accurate predictions. Moreover, this detection is quantified in a way that is portable. Given this portability, the gains in going from the 1PL to the 2PL can be compared directly to subsequent gains we observe from other kinds of modeling innovation.

We now make such a comparison by looking at gains associated with going from the 2PL to the 3PL. Consider the red points in Figure 1, which show IMV(2PL,3PL). In contrast to the results shown for IMV(1PL,2PL), the 3PL never provides much additional value relative to the 2PL irrespective of the value of C. Average IMV values are never greater than 0.001. While the evaluation of the IMV(1PL,2PL) values suggested that the IMV was sensitive to that modeling change, consideration of IMV(2PL,3PL) suggest that the IMV is able to detect when adding model complexity does not improve predictions of new data. In SI-S2.2 we show that the IMV(2PL,3PL) values remain small even when guessing is more pronounced. This is due to previously identified problems related to the identification of 3PL model parameters (Maris & Bechger, 2009; Haberman, 2005; von Davier, 2009) and is a topic we return to in Section 5.

In the above, we consider comparisons between different IRT models. One advantage of the IMV is that it allows comparison across very different types of models that do not need to be from the same family of models. We illustrate this by also considering non-IRT mechanisms for generating p_0 or p_1; for example, we examine IMV($\bar{x_j}$,1PL), where $\bar{x_j} = \sum_i x_{ij}/N$ represents the response probability as calculated by the in-sample proportion of correct responses for each item (i.e., the classical item p-value). Note that this probability is constant across persons within each item and describes the value of the 1PL versus predictions that do not account for between-person differences in ability. We can similarly consider IMV($\bar{x}, \bar{x_j}; x^\star$), where $\bar{x} = \sum_{i,j} x_{ij}/N$ represents the mean response across all items and people. This quantity describes the value of predictions that account for item-level differences in difficulty as compared to a universal prediction based on overall difficulty alone.

The average IMV($\bar{x_j}$,1PL;$x^\star$) across all iterations in Figure 1 was 0.1. Similarly, the average IMV($\bar{x}, \bar{x_j}; x^\star$) across all iterations was 0.26. Note that the IMV values from this simulation are maximal in the sense that they would be smaller if $\mathbb{E}(\theta_i - b_j) \neq 0$; we illustrate this fact in SI-S2.1. These IMVs indicate that the gains associated with the inclusion of item-level variation in the discrimination parameter relative to difficulty alone are an order of magnitude less valuable than allowing variation in p_{ij} after considering both item- and person-parameters.

3.2 Evaluation of model-based predictions versus truth and a consideration of overfitting

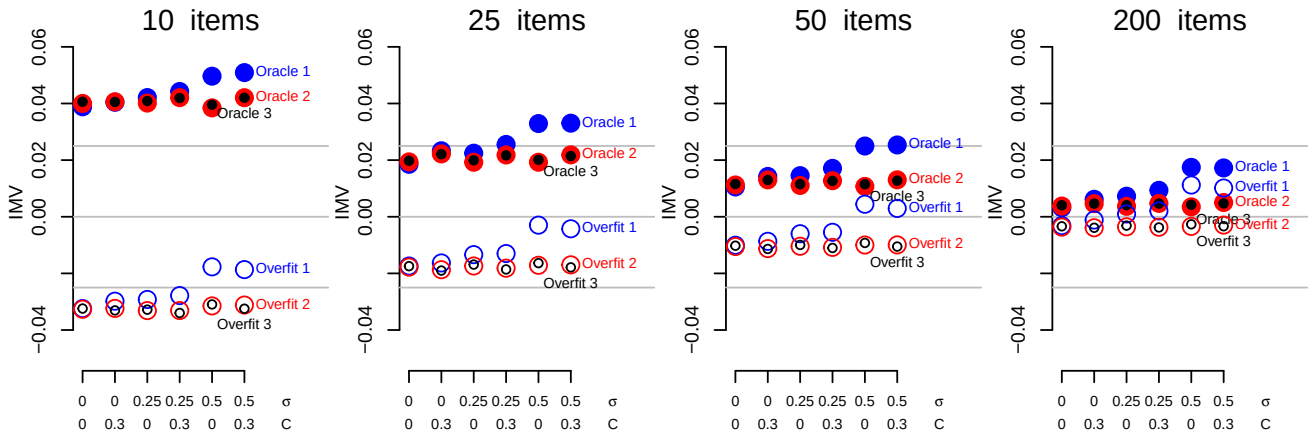
We now evaluate the balance between prediction accuracy and overfitting. To do this, we compare model-based predictions to the true probabilities. Predictions from in-sample data suffer from overfitting which we show by computing IMV values tailored to index overfitting using in-sample data. Predictions on out-of-sample data do not have this problem and are used to quantify the value that absolute truth (i.e., predictions from an oracle) have versus model-based predictions.

Formally, we compute Oracle values as IMV($\widehat{p_{ij}}, p_{ij}; x^\star$); that is, we are asking about the value that would be associated with knowing the true p_{ij} value relative to our estimate $\widehat{p_{ij}}$ for out-of-sample data. Similarly, Overfit is defined as IMV($\widehat{p_{ij}}, p_{ij}; x$). As with the Oracle, we are again asking about the value associated with knowing the true p_{ij} quantities relative to the estimates based on an IRT-model. The key distinction is that for the Overfit we are computing IMVs based on in-sample data (x rather than $x^\star$). If the Overfit value is below zero, this implies that the estimates are better predictors than the truth (a clear indication of overfitting). We will use $\widehat{p_{ij}}$ derived from the 1PL, 2PL, and 3PL models. We emphasize that these Oracle and Overfit values, as compared to the quantities such as IMV(1PL,2PL) considered in Figure 1, are only available given that we are working with simulated data and thus know truth (i.e., p_{ij}).

The results are shown in Figure 2. Consider the 1PL: for these estimates, there is increasing value in the Oracle as σ and C increase. In contrast, the Oracle is positive but relatively constant across σ and C for the 2PL and 3PL thus suggesting that it is the ability to fit the changing discrimination parameters that is resulting in valuable gains in estimates of item-level response probabilities. This dovetails with the previous observation regarding the fact that IMV(2PL,3PL) is very near zero irrespective of C. The fact that, for a given value of C, we see no change in the Oracle for the 2PL across values of σ indicates that it is the flexibility associated with fitting varying discrimination parameters, not the level of variation in those parameters, that leads to predictive value from the 2PL relative to the truth (of course, IMV(1PL,2PL) increases as a function of σ, see Figure 1).[6] Note that the value of the Oracle IMV when $\sigma = C = 0$ is similar across all three IRT models (i.e., the points in each panel overlap) but depends on the number of items (i.e., it is near 0.02 for 25 items but less than 0.01 for

[6] Ancillary analyses suggest that we do observe systematic variation in the 2PL Oracle across values of σ if the sample size is smaller.

Figure 2: Oracle and Overfit values computed for simulations in Figure 1. Solid dots represent Oracle values where the number indicates the IRT model (e.g., 3 indicates the 3PL). Hollow dots indicate Overfit values.



200 items); we further discuss this dependency on sample size below.

Turning to the Overfit values, we would generally expect better prediction when we have access to the true p_{ij} values as compared to the estimated values $\widehat{p_{ij}}$. However, note that the Overfit values in Figure 2 are generally negative. This confirms that the $\widehat{p_{ij}}$ values are overly tailored to x. The penalty is relatively constant for the 2PL and 3PL as a function of σ and C, but the magnitude of the penalty depends on the number of items and is nearly zero for the largest number of items (i.e., estimates of $\widehat{p_{ij}}$ are nearly as good as p_{ij}). There are interesting features of the 1PL Overfit estimates. Consider the 25 item case. When the true model is effectively the 1PL (i.e., $\sigma = C = 0$), all three IRT approaches see an expected Overfit IMV of nearly -0.02. However, as σ and C increase, the penalty associated with the 1PL model declines to nearly 0. A similar story holds as we increase the number of items and, in fact, the Overfit associated with the 1PL is actually positive when σ and C are relatively large, suggesting that the true p_{ij} values are more predictive than the estimated $\widehat{p_{ij}}$ values. This indicates that the 1PL is more robust to Overfitting than the 2PL and 3PL when the data generating model is relatively complex but, of course, this comes at the cost of being overly restrictive in predicting outcomes when the 1PL is not the correct model (especially when $\sigma > 0$).

3.3 Prediction accuracy as a function of sample size

When appropriate models are applied, larger samples should allow for more accurate estimation of model parameters. We thus use the IMV to index changes in predictive value as a function of sample size. We extend the above simulations to allow for varying numbers of respondents (up to 10,000). Results can be found in the supplementary materials (SI-S2.5). We demonstrate that the (IMV-derived) costs of misfit are not strongly sensitive to sample size except for the case of the IMV(1PL,2PL) if the data generating model is the 2PL or 3PL (SI-S2.5) which doubles from around 0.005 for 100 respondents to over 0.01 when there are several thousand respondents. These results are consistent with the notion that more respondents allow for more accurate estimates of slope

parameters but do not translate into more accurate guessing parameters. Note that we are not arguing that the IMV's value is sensitive to sample size in a way that invalidates comparisons; rather, predictions are improved (in some cases) with larger samples and these improved predictions lead to larger IMVs.

In the SI, we further explore the sensitivity of the Oracle values to sample size (SI-S2.6); we note two key findings. First, the Oracle values behave as expected as a function of sample size (i.e., they move towards zero for larger number of respondents and a fixed number of items). Second, there is a lower bound on the value of the Oracle for a fixed number of items. That is, increasingly large samples do not further generate value for a fixed number of items. This is because the $\widehat{p_{ij}}$ estimates cease to become more accurate given that the precision of ability estimates is limited by the sample size of items.

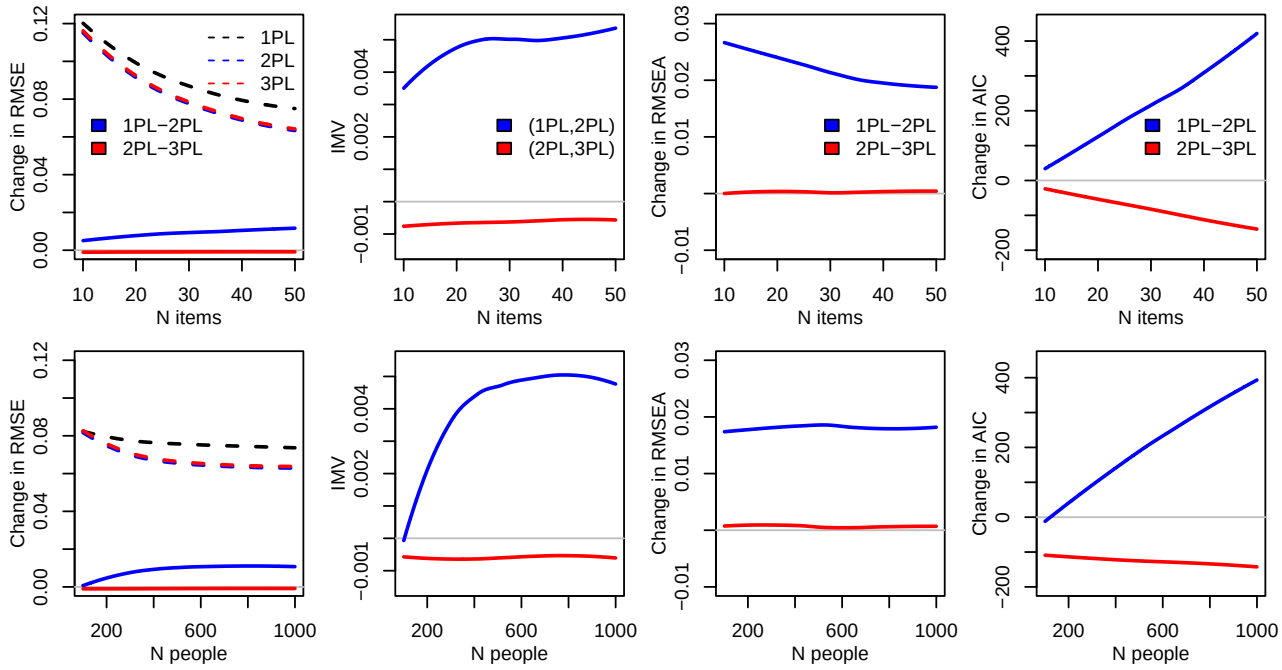
3.4 A comparison of the IMV to alternative metrics

If interest is in a simple up/down decision about one model or another, we anticipate that commonly used metrics will provide similar information under many common scenarios. This is desirable; the IMV will typically provide the same information as other metrics if interest is solely in that decision (see simulations in Domingue et al., 2021). However, the IMV provides qualitatively different information if the focus is on the questions of 'how much better?' rather than strictly 'which is better?'. We offer an illustration of this point focusing on commonly used indices in IRT settings with an emphasis on both the portability of the IMV and the fact that it is indexing prediction quality. These results, when joined with those from earlier work (Domingue et al., 2021), suggest that the IMV offers novel information compared to conventional metrics. We consider comparisons to the AIC (Akaike, 1973)—specifically, the difference in the AIC (Burnham & Anderson, 2004) between sequential models (i.e., $\text{AIC}_{\text{1PL}} - \text{AIC}_{\text{2PL}}$, $\text{AIC}_{\text{2PL}} - \text{AIC}_{\text{3PL}}$)—and the RMSEA based on the M2 statistic (Maydeu-Olivares, Cai, & Hernández, 2011). The AIC is widely used for model selection and involves a penalty based on the number of estimated parameters. Note that AIC does not depend on number of estimated person abilities. The RMSEA considers parsimony by capturing the amount of model misspecification per degree of freedom (Browne & Cudeck, 1992); for comparisons to the other indices, we consider differences in RMSEA values. We also take advantage of the simulated setting to consider RMSEs between true and estimated probabilities. While it would not be useful in empirical settings where truth is not known, the behavior of the RMSEs offers a valuable benchmark here in that it helps to calibrate our understanding of the other metrics. In particular, we use the RMSE as our benchmark for gauging changes in predictive accuracy.[7]

We base simulations on the 2PL (Eqn 6 with $c_j = 0$). We sample $b_j \sim \text{Normal}(0, 1)$ and $\log a_j \sim \text{Normal}(0, 0.3^2)$. We focus on the implications of sample size. Results for this first simulation study are shown in Figure 3; we separately vary the number of items (top; $N \sim \text{Unif}(10, 50)$) and the number of respondents (bottom;

[7]Note that the RMSE would not, in general, be portable across simulation studies given the dependence on $\mathbb{E}(\theta - b_j)$. However, in the simulation study here we do not vary $\mathbb{E}(\theta - b_j)$ and thus are focusing on values of the RMSE that are directly comparable.

Figure 3: 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 (shown as different colors). Data are generated via the 2PL based on different numbers of items (top; 1000 respondents) or different numbers of people (bottom; 50 items). Solid lines indicate comparisons; in the first row, the dashed lines indicate raw RMSEs for the 3 models. Results are based on LOESS smoothing for 1000 choices of the component of sample size being varied (top, items; bottom, people).



$N \sim \text{Unif}(100, 1000)$). Consider first the root mean squared errors (RMSE; in the left panel), which contrasts estimates from a given model with true probabilities used to generate responses; the RMSE would not be available in practice but is useful here given that it allows us to benchmark the behavior of the various metrics to the RMSE's comparisons of estimates to truth. The RMSE is smallest in absolute terms for the 2PL—given that the 2PL was used to generate the data the 3PL is overfit to the in-sample data (see also results from Figure 1 when $\sigma = 0$)—with the 1PL estimates being worse in larger samples. Note also that there is some increase in the difference in the RMSE for the intial growth in sample (either items or persons) but this growth seems to level out (especially as a function of the number of people) for larger sample sizes. A metric that is sensitive to improvements in prediction accuracy should behave similarly.

Turning to the metrics that are computed based on estimated quantities (rather than true probabilities), the IMV(1PL,2PL) increases as sample size increases while the IMV(2PL,3PL) is negative but quite small and fairly insensitive to changes in sample size. Both of these results are anticipated given the RMSE results. The IMV is sensitive in that increases in sample size lead to improvements in predictive accuracy. Suppose we go from 200 to 500 respondents being used to estimate model parameters, this increase results in increases in accuracy for the 2PL relative to the 1PL that we can observe in the RMSE and we similarly observe increases in the IMV. Larger

sample sizes (i.e., going from 500 to 1000) do not yield tangible differences in predictive accuracy (again see the RMSE) and the IMV is similarly flat.

For the RMSEA and AIC values, we focus interpretation on the areas of difference. The RMSEA does little to capture the divergence in 1PL and 2PL/3PL predictions as the number of respondents increases (the RMSEA also misidentifies the generating model in SI-S2.7). The AIC's sensitivity to the number of estimated item parameters is apparent in the upper right panel; rather than level out as do the RMSE and IMV values the AIC differences are largely linear. These panels help to illustrate the issue of portability associated with the AIC. The RMSE values, for example, are always comparable for the 2PL and 3PL irrespective of sample size. However, the AIC differences comparing the 2PL and 3PL results vary from near 0 to nearly -200 depending on the relevant sample sizes. In contrast to what we obsered with the IMV, the RMSEA is not sensitive to changes in predictive accuracy (note the flatness of the curves in the bottom panel) and the AIC is not portable (the AIC grows linearly as a function of the sample size irrespective of the predictive gains suggested by the RMSE). In SI-S2.7, we further illustrate these points of difference between the IMV and AIC/RMSEA by focusing on variation in μ when $b_j \sim \text{Normal}(\mu, 1^2)$.

3.5 Synthesizing the simulation results

We synthesize IMV values from the simulation studies in Table 1. These provide generic guidance about the kind of increase in predictive performance that comes from different modeling innovations. We offer them for the purposes of helping to develop intuition about the degree to which modeling choices affect predictive accuracy. When the data generating model is a 1PL, adoption of item-level variation in prediction is incredibly valuable, $\text{IMV}(\bar{x}, \bar{x_j}) = 0.3$, relative to prediction based on the overall p-value alone. Prediction based on the 1PL leads to $\text{IMV}(\bar{x_j}, \text{1PL}) = 0.1$; incorporation of person-level variation is, not surprisingly, quite useful for improving predictions even after item-level variation has been included.

Turning now to data generated from more complex models, prediction based on the 2PL rather than the 1PL is an order of magnitude less valuable, $\text{IMV}(\text{1PL}, \text{2PL}) = 0.01$, than $\text{IMV}(\bar{x_j}, \text{1PL})$. This value depends on a choice of σ. Here, we focus on results for $\sigma = 0.5$ which corresponds to a_j parameters that whose 10% and 90% percentiles range from 0.53 to 1.88; different choices of σ lead to different IMV values (see Figure 1), we return to this point in our discussion of empirical results. As an additional point of comparison, we can compute the IMV based on different approaches to generating ability estimates. After computing both MLE and EAP estimates, we can compute IMV(MLE,EAP), see SI-S2.3. Predictions based on the EAP versus the MLE are nearly an order of magnitude less valuable still, IMV(MLE,EAP)=0.002. Transitioning from the 2PL to the 3PL as the generative model, recovery via the 3PL has an IMV much smaller than the switch in ability estimation methods, IMV(2PL,3PL)=0.0005.

To emphasize the portability of the IMV, we can also compare these values to those generated in non-IRT

Table 1: Approximate expected IMVs for different modeling scenarios. Results based on simulated item response data for 50 items and 1000 respondents using the appropriate generating model.

IMV Comparison	Value	Generating Model(s)	Source
IMV($\bar{x}$,$\bar{x_j}$)	0.3	1PL	SI-S2.1[a]
IMV($\bar{x_j}$,1PL)	0.1	1PL	SI-S2.1[a]
IMV(1PL,2PL)	0.01	3PL	Figure 1[b]
IMV(MLE,EAP)	0.002	3PL	SI-S2.3
IMV(2PL,3PL)	0.0005	3PL	Figure 1[b]

[a] Data generated via Eqn 6 with $a_j = 0$, $c_j = 0$, $\mathbb{E}(b) = 0$, and $\text{Var}(b) = 1$.
[b] Data generated via Eqn 6 with $\sigma = 0.5$ and $C = 0.3$.

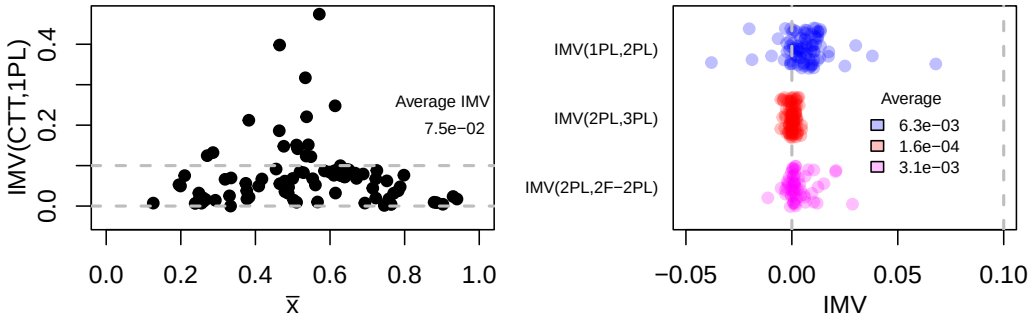
settings (Domingue et al., 2021). The value of allowing for item-level variation in predictions (IMV$(\bar{x}, \bar{x_j}) = 0.3$) is similar to the value of demographics in predicting the political affiliation of US adults in 1991. Moving to the 1PL (IMV$(\bar{x_j}, 1\text{PL}) = 0.1$) is akin to the value that self-reported symptoms (e.g., loss of taste) provided in predicting COVID infections early in the outbreak. Relative to the 1PL, the 2PL provided predictive value on the order of what age and sex provide in predicting high blood pressure amongst respondents near 63y of age. Such comparisons are useful in helping us understand the general utility of modeling improvements made in IRT by allowing us to contrast them with the predictive gains observed in other contexts.

We believe the values in Table 1 have implications for application. For example, there is a relatively large literature on the sample size requirements of the 3PL (see discussion in Feuerstahler, 2020). In our view, the difference between the 2PL and 3PL is fairly negligible in terms of the value offered by the predictions *even when the 3PL is the true data generating model*. This is not to say that estimates from the 3PL can never help to identify items that have issues associated with guessing or that such information would not be valuable; rather, we are asserting that, if interest is in the quality of the resulting predictions, the difference between the 2PL and 3PL is likely negligible.

4 The IMV in empirical data

In order to evaluate how different models improve prediction of item responses in practice, we apply the IMV approach to 89 dichotomously scored item response datasets taken from the IRW (Domingue & Kanopka, 2023). These data span a range of cognitive and affective tasks. We consider the broad range of data for the purpose of both illustrating common trends related to variation in prediction quality and also studying interesting exceptions revealed by the IMV values. The median dataset has 1500 respondents (range 118–10000) and 36 items (range 4–529). For a description of each dataset and its associated results, please see SI-S3.

Figure 4: IMVs computed using different models with 89 empirical datasets. Gray lines are similarly placed in each figure to emphasize comparability across results (and average IMVs for each approach are also shown). Left: IMV(CTT,1PL) as a function of the average response $\bar{x}$ in a given dataset. Right: IMVs contrasting a range of IRT models.



We focus on analysis of IMVs computed for various modelling approaches and use the quantities in Table 1 as a means of understanding the relative magnitude of these values. We use the same analysis pipeline as with simulated data with a few exceptions. To minimize the computational costs, we consider a random sample of 10,000 respondents in datasets with large numbers of respondents. To compute the IMV, we implement the "missing response" paradigm (Stenhaug & Domingue, 2022) by splitting the response-level data at random into $k = 4$ folds.[8] For a given fold, we use all *responses not in that fold* to produce estimates of ability and item parameters, and then combine those to generate predictions for the *responses in the fold*. The same priors as in the simulations were used here for estimation. IMVs are computed based on those predictions; we take the average IMV across the folds.

We begin by comparing the predictions from the 1PL model to those that use simply the item-level mean (i.e., $\bar{x_j}$). A visualization of the IMVs as a function of the average response $\bar{x}$ is shown in Figure 4; correlations between dataset descriptive statistics and IMV results are shown in Table 2. The average value of IMV($\bar{x_j}$,1PL) was 0.075; note that this is less than but in proximity to the 0.1 value from simulation studies (e.g., Table 1 when the 1PL was the true model). There are some extreme values; the largest value (IMV($\bar{x_j}$,1PL)=0.47) is for data from a four-item attitudinal survey regarding abortion (see Rizopoulos, 2006).

We emphasize a few additional points related to these IMV($\bar{x_j}$,1PL) values. First, as expected, IMVs are generally larger for datasets that have items with correct response rates near 50% ($r = -0.42$, see Table 2). This is due to the fact that the level of uncertainty for the Bernoulli random variable varies with prevalence; the IMV is related to uncertainty and there is simply less uncertainty when prevalences are far from 0.5 for models to explain. That said, for a given average response level, there is still variation in the IMV thus indicating that the IMV is sensitive to prediction quality on top of overall difficulty. Second, the IMVs are effectively independent of the number of people (correlation of -0.07) and only weakly associated with the number of items (-0.22); we offer a

[8]We conduct a small simulation study regarding the choice of k, see SI-S3.2. Resulting IMVs are relatively insensitive to the particular choice of k.

Table 2: Correlations between IMV values (for 1PL, 2PL, 3PL, and 2F-2PL models) and key dataset descriptives (numbers of people and items; $|.5 - \bar{x}|$ where $\bar{x}$ is the average response for a dataset) for empirical analyses.

	1PL	2PL	3PL	2F-2PL
N people	-0.07	0.16	0.43	-0.32
N items[a]	-0.22	0.13	0.06	0.31
$\|.5 - \bar{x}\|$	-0.42	-0.25	-0.21	-0.07
1PL	1.00	-0.39	-0.03	-0.15
2PL		1.00	0.10	0.48
3PL			1.00	-0.06

[a] One dataset had over 500 items; correlations with number of items is computed with this dataset removed.

potential explanation for this negative correlation with the number of items below. Third, to indicate the flexibility of the IMV, we also compare estimates from the 1PL model to an alternative baseline: that of the Guttman model (Guttman, 1950).[9] The expected payoffs in this case are quite large (average IMV(Guttman,1PL)= 0.53) which is consistent with arguments regarding the utility of probabilistic item response models (Sijtsma, 2012).

Turning to more complex models beginning with the 2PL, we observe an average IMV(1PL,2PL)=0.006, an order of magnitude smaller than the mean IMV($\bar{x_j}$,1PL). The IMV associated with moving from the 1PL to the 2PL was somewhat larger in simulation studies (IMV(1PL,2PL)=0.01) than the average here but recall that this quantity depended on a choice of σ; we view the proximity of these empirical results to the value observed in simulations as supportive of the IMV values associated with $\sigma = 0.5$ that we focus on in Table 1. There is variation in these IMV(1PL,2PL) quantities (max of 0.068 for data from toddlers on balance-problem items; Van Maanen, Been, & Sijtsma, 1989); note that there was also significant variability in the simulation-based IMV(1PL,2PL) values of Figure 1. The average IMV(2PL,3PL) is 0.00016 with a max of 0.0046 (data from a literacy intervention; Gilbert, Kim, & Miratrix, 2023); these results suggest weak improvements in predictive value for the 3PL relative to the 2PL as was observed in simulation studies.

As a contrast to the 3PL results, we also consider a fully exploratory two factor 2PL (2F-2PL). Data restrictions led to analysis in only 60 datasets.[10] The average IMV(2PL,2F-2PL) was 0.0031; this is somewhat larger than

[9] Specifically, we use the person-level ability estimates θ_i and the 1PL difficulty estimates δ_j and set the probability of response where $\theta_i > b_j$ as 0.99 and 0.01 otherwise (note that we cannot use 1 and 0 respectively given that these would lead to malformed likelihoods).

[10] We considered multidimensional analysis of a subset of the 89 empirical datasets. We required that the dataset contain observations from at least 500 respondents and used a random sample of 25,000 respondents rather than 10,000 respondents for larger datasets. We also omit results for datasets wherein either the estimation algorithm did not converge or the estimated abilities were effectively identical (a mean absolute deviation between abilities of less than 0.001). These restrictions left us with results for 60 datasets.

the IMV(2PL,3PL) results but there was also more variability in the right tail. The maximal value was 0.029 (data from a personality inventory; Eysenck & Eysenck, 1968); indeed, many of the cases wherein the 2F-2PL offered large preditive increases were based on personality inventories that are conventionally assumed to be multidimensional. We thus argue that the 2F-2PL is a modeling innovation that is able to produce tangible gains in predictive value in some tailored cases; the 3PL only provides—at best—very weak predictive value. When combined with the simulation evidence (e.g., Figure SI-S2.2), we are pessimistic about the utility of fitting the 3PL.

We offer two additional notes about the multidimensional results. First, as context for the larger IMVs observed for the IMV(2PL,2F-2PL), we considered a simulation study (see SI-S2.8); IMVs greater than 0.02 can be obtained even when we simulate data with fairly strongly correlated latent factors ($\rho > 0.5$). Second, note that, as one may expect, the correlation between the IMV(2PL,2F-2PL) and the number of items was positive ($r = 0.31$) which may be one component of the observed negative correlation between IMV($\bar{x}_j$,1PL) and the number of items.

5 Discussion

A great volume of psychometric research is concerned with adjudicating between different modeling approaches. A variety of approaches are available for making such decisions. In our view, these approaches have many shared weaknesses. In particular, we are concerned about a lack of portability across settings that leads to an impoverished intuition about the predictive gains associated with modeling choices amongst applied researchers. The IMV is a different approach that quantifies the predictive value encoded in one model relative to another. It quantifies the gain in prediction in the form of expected winnings in bets due to the side information encoded in the focal model as compared to some baseline. In contrast to most existing fit indices, the IMV is a predictive index that assesses a model's success at out-of-sample prediction. As such, it is well suited to address an increasing interest in prediction (even when explanation is the ultimate goal; Yarkoni & Westfall, 2017) and is highly portable in that its values can be meaningfully compared across a variety of contexts. In this paper, we describe how the IMV can be used with IRT models of dichotomous item responses.

We described a sequence of simulation studies that are collectively meant to demonstrate the utility of the IMV as a means of understanding the functioning of IRT models. We studied the IMV as a tool for understanding misfit, for example. Of particular interest is the observation that IMV(2PL,3PL) tends to be near zero as the 2PL can effectively approximate guessing via adjustment to difficulty and discrimination parameters in a way that makes the resulting out-of-sample 2PL estimates of $\widehat{p_{ij}}$ highly comparable to those produced by the 3PL. Results here are similar to others (e.g., Stenhaug & Domingue, 2022) in suggesting that the 3PL might have limited utility in many settings.

We also use the inherently comparative nature of the IMV to introduce the Oracle and the Overfit values. We use these values to illustrate a few pertinent facts. When the 1PL model is used to simulate data, there is no value associated with fitting more complex models. This is expected since the innovations of the 2PL and 3PL are not necessary. Note that, even in this simple case, the IMV of truth relative to estimates (i.e., the Oracle) does not decline to zero as a function of the number of respondents alone; the number of items also needs to get increasingly large. This is a useful reminder regarding the limited utility of having ever more respondents. Even when the 2PL is used to generate data, estimates from the 2PL are overfit to the data in a way that is not true of 1PL estimates (i.e., true response-level probabilities are better predictors of new data than 1PL-based estimates but worse than 2PL-based estimates). In this sense the 1PL-based estimates are lower variance but higher bias in the sense of the bias-variance trade-off (Doroudi, 2020).

We also compared the IMV to alternative metrics. The IMV was observed to be reflective of variation in the RMSE between true and estimated probabilities of responses as a function of various quantities manipulated in simulation studies in a way that made it distinctive as compared to the AIC and RMSEA. We also emphasize the ease of interpretation of the IMV. While there is guidance on generic interpretations of the other indices (e.g., Browne & Cudeck, 1992 for the RMSEA and Burnham & Anderson, 2004 for the AIC differences), the interpretation of the IMV is aided by the quantities shown in Table 1. We used simulations to offer context to the improvements in prediction associated with different modeling innovations that, we believe, could be highly useful in future work. Table 1 describe the level of predictive power that we should expect under known conditions. For example, we should anticipate IMV(1PL,2PL) approaching 0.01 when there is in fact substantial variation in the discrimination parameters (recall that we simulate $\log a_j \sim \text{Normal}(0, \sigma^2)$ with $\sigma = 0.5$). Future work with the 2PL can still be informed by this rough benchmark (and, of course, more precise benchmarks can be obtained; e.g., McNeish & Wolf, 2021). Finally, note also the flexibility of the IMV. We use it here to study the effects of sample size, priors, and estimation algorithms on prediction quality; this flexibility is a vital component of the IMV's appeal.

In our view, the evidence from simulation and empirical work suggests that the IMV is a useful new tool for understanding the performance of IRT models. The IMV approach provides a complementary perspective based on the level of predictive difference across models rather than attempts to ascertain the true model; in particular, values from application of different IRT models in empirical settings can be compared to the quantities in Table 1 so as to indicate whether the performance of a model in a given data context is unique or as-expected. The IMV's focus on how models predict new data is important. We agree with others (Yarkoni & Westfall, 2017) that prediction is relevant even when the goal is to identify mechanisms.

The simplicity of the IMV—it merely requires predictions of responses generated by any mechanism—suggests that it could be further used in other settings. For example, the IMV could be used to understand the behavior of specific items. There are also possibilities of further extending the IMV to deal with non-dichotous responses.

While there are complexities associated with such extensions, advances on this front would be of potential utility as they might allow for straightforward comparison of the performance of IRT models across both dichotomous and polytomous items. This simplicity comes at one potential cost, however. The IMV's reliance on cross-validation may necessitate somewhat larger samples; future work can focus on the implications of using this method with smaller samples.

In the future, the IMV could be used to further quantify the degree to which modeling innovations provide value in predicting new data relative to conventional alternatives. Innovations in psychometric techniques are welcome, even ones that produce relatively marginal predictive improvements. However, it is our view that a firmer foundation for future development of psychometric models would include a generalizable tool for understanding the magnitudes of predictive improvement offered by a given innovation. The IMV is such a tool.

References

Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, *60*(2), 255–265.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological methods & research*, *21*(2), 230–258.

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

Cai, L., Chung, S. W., & Lee, T. (2021). Incremental model fit assessment in the case of categorical data: Tucker–lewis index for item response theory modeling. *Prevention Science*, 1–12.

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

Craven, P., & Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische mathematik*, *31*(4), 377–403.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* ERIC.

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). Intermodel vigorish (imv): A novel approach for quantifying predictive accuracy when outcomes are binary. Retrieved from `https://osf.io/gu3ap/`

Doroudi, S. (2020). The bias-variance tradeoff: How data science can inform educational debates. *AERA Open*, *6*(4), 2332858420977208.

Eysenck, H. J., & Eysenck, S. B. (1968). Eysenck personality inventory. *Journal of Clinical Psychology*.

Feuerstahler, L. M. (2020). Metric stability in item response models. *Multivariate Behavioral Research*, 1–18.

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.

Guttman, L. (1950). The basis for scalogram analysis. *Measurement and prediction*, 60–90.

Haberman, S. J. (2005). Identifiability of parameters in item response models with unconstrained ability distributions. *ETS Research Report Series*, *2005*(2), i–22.

Haberman, S. J., Sinharay, S., & Lee, Y.-H. (2011). Statistical procedures to evaluate quality of scale anchoring. *ETS Research Report Series*, *2011*(1), i–20.

Han, Y., Zhang, J., Jiang, Z., & Shi, D. (2022). Is the area under curve appropriate for evaluating the fit of psychometric models? *Educational and Psychological Measurement*, 00131644221098182.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, *143*(1), 29–36.

Haslbeck, J., & van Bork, R. (2024). Estimating the number of factors in exploratory factor analysis via out-of-sample prediction errors. *Psychological Methods*, *29*(1), 48–64.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., ... others (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kang, T., & Cohen, A. S. (2007). Irt model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331–358.

Köhler, C., Robitzsch, A., & Hartig, J. (2020). A bias-corrected rmsd item fit statistic: An evaluation and comparison to alternatives. *Journal of Educational and Behavioral Statistics*, *45*(3), 251–273.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. IAP.

Maris, G., & Bechger, T. (2009). On interpreting the model parameters for the three parameter logistic model. *Measurement*, *7*(2), 75–88.

Mavridis, D., Moustaki, I., & Knott, M. (2007). Goodness-of-fit measures for latent variable models for binary data. In *Handbook of latent variable and related models* (pp. 135–161). Elsevier.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 71–101.

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.

Maydeu-Olivares, A., & Garcia-Forero, C. (2010). Goodness-of-fit testing. *International encyclopedia of education*, *7*(1), 190–196.

Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in

2 n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*(471), 1009–1020.

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*.

Rahal, C., Verhagen, M., & Kirk, D. (2022). The rise of machine learning in the academic social sciences. *AI & Society*.

Reddy, S., Labutov, I., Banerjee, S., & Joachims, T. (2016). Unbounded human learning: Optimal scheduling for spaced repetition. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1815–1824).

Rizopoulos, D. (2006). ltm: An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. Retrieved from `https://doi.org/10.18637/jss.v017.i05`

Savalei, V., Brace, J., & Fouladi, R. T. (2021, May). *We need to change how we compute rmsea for nested model comparisons in structural equation modeling.* PsyArXiv. Retrieved from `psyarxiv.com/wprg8` doi: 10.31234/osf.io/wprg8

Savcisens, G., Eliassi-Rad, T., Hansen, L. K., Mortensen, L. H., Lilleholt, L., Rogers, A., ... Lehmann, S. (2023). Using sequences of life-events to predict human lives. *Nature Computational Science*, 1–14.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.

Shmueli, G. (2010). To explain or to predict? *Statistical science*, *25*(3), 289–310.

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology*, *22*(6), 786–809.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, *64*(4), 583–639.

Stenhaug, B., & Domingue, B. (2022). Predictive fit metrics for item response models. *Applied Psychological Measurement*.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 44–47.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589–617.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2006). 21 assessing the fit of item response theory models. *Handbook of statistics*, *26*, 683–718.

Van der Linden, W. J. (2017a). *Handbook of item response theory: Volume 2: Statistical tools.* CRC Press.

Van der Linden, W. J. (2017b). *Handbook of item response theory: Volume 3: Applications.* CRC press.

Van Maanen, L., Been, P., & Sijtsma, K. (1989). Problem solving strategies and the linear logistic test model. In *Mathematical psychology in progress* (pp. 267–287). Springer.

Verhagen, M. D. (2022). A pragmatist's guide to using prediction in the social sciences. *Socius*, *8*, 23780231221081702.

von Davier, M. (2009). Is there need for the 3pl model? guess what? *Measurement: Interdisciplinary Research and Perspectives*, *27*.

Wagenmakers, E.-J., & Farrell, S. (2004). Aic model selection using akaike weights. *Psychonomic bulletin & review*, *11*(1), 192–196.

Wainer, H. (2016). Discussion of david thissen's bad questions: An essay involving item response theory. *Journal of Educational and Behavioral Statistics*, *41*(1), 100–103.

Watts, D. J. (2014). Common sense and sociological explanations. *American Journal of Sociology*, *120*(2), 313–351.

Watts, D. J. (2017). Should social science be more solution-oriented? *Nature Human Behaviour*, *1*(1), 1–5.

Watts, D. J., Beck, E. D., Bienenstock, E. J., Bowers, J., Frank, A., Grubesic, A., ... Salganik, M. (2018). Explanation, prediction, and causality: Three sides of the same coin?

Wolfram, T., Tropf, F. C., & Rahal, C. (2022, May). *Short essays written during childhood predict cognition and educational attainment close to or better than expert assessment.* SocArXiv. Retrieved from `osf.io/preprints/socarxiv/a8ht9` doi: 10.31235/osf.io/a8ht9

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th ed.). Cengage Learning.

Wu, M., & Adams, R. J. (2013). Properties of rasch residual fit statistics. *Journal of Applied Measurement*.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

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.[11]

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
}
```

[11] 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 et al., 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

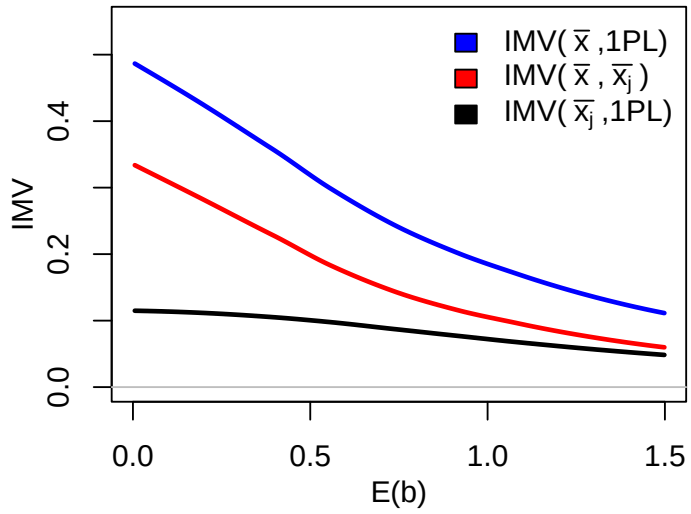
```
set.seed(170301)
library(mirt); library(redivis); library(irw)
dataset <- redivis::user("datapages")$
    dataset("item_response_warehouse",version='v2.0')
df <- dataset$table("content_literacy_intervention")$to_data_frame()
df$item<-paste("item_",df$item,sep='')
ntimes<-4
df$gr<-sample(1:ntimes,nrow(df),replace=TRUE)
omega<-numeric()
for (i in 1:ntimes) {
    x<-df
    x$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.



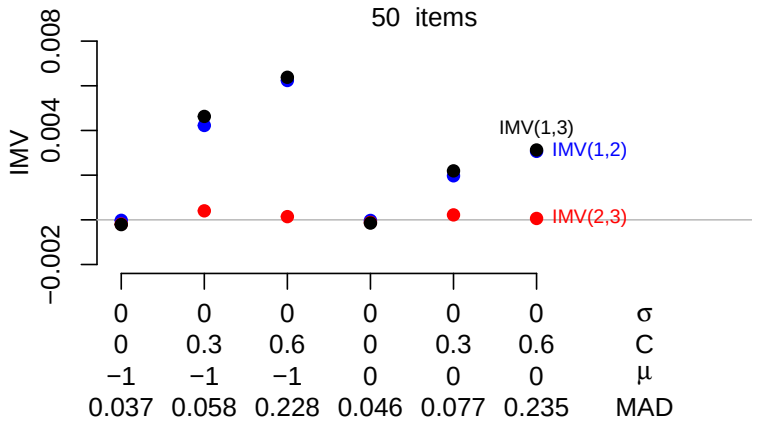
S2.2 Further analysis of the 3PL

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.



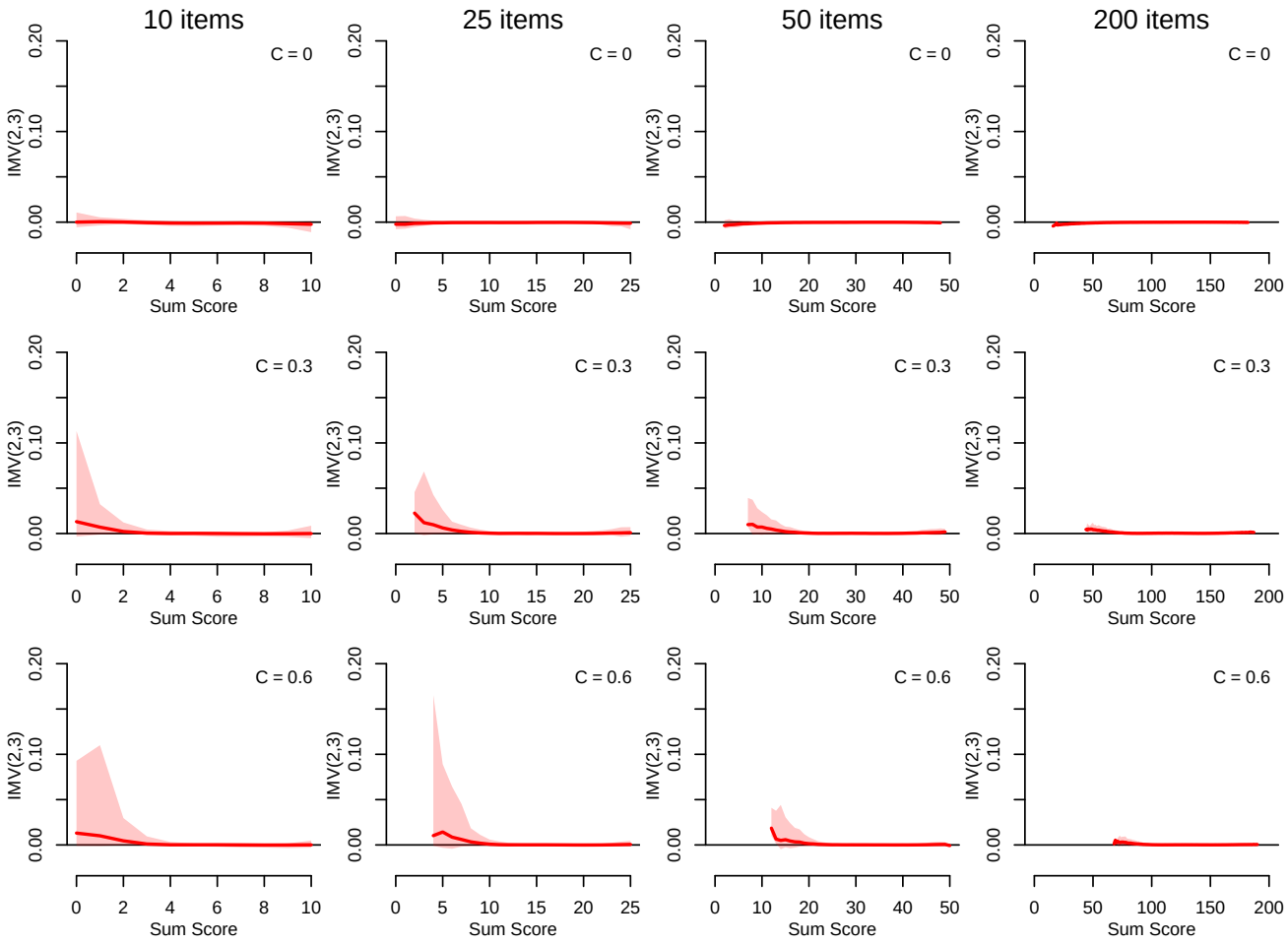
S2.3 The EAP versus the MLE

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

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.



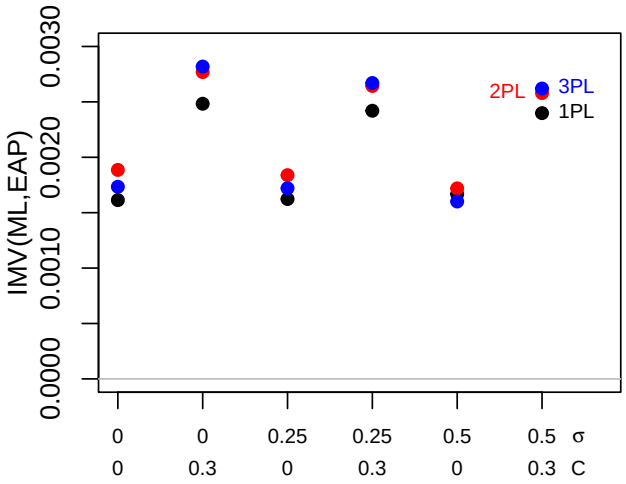
parameters of LogNormal(m, s).[12] We let $m \in \{0, 0.2\}$ and sampled 250 values for $s \sim \text{Unif}(0.05, 1)$. We consider sample sizes of 1000 respondents and 50 items.

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 Unif$(0, C)$ where

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

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.



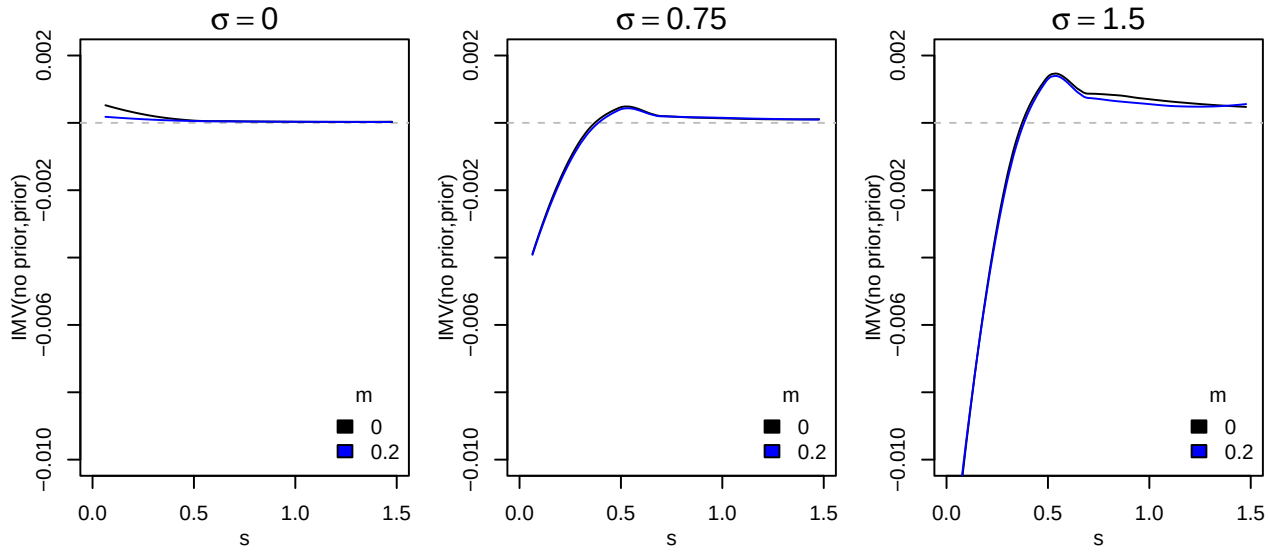
we vary C in simulation. For estimation, we use priors of the form Beta(2,β). Interest is in the IMV for values of β relative to $\beta = 17$ which we use in other analyses. We show the variation in the prior as a function of the choice of β in the left-hand panel of Figure S6. We emphasize here that 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.

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 2PL continues to generate a high IMV

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).



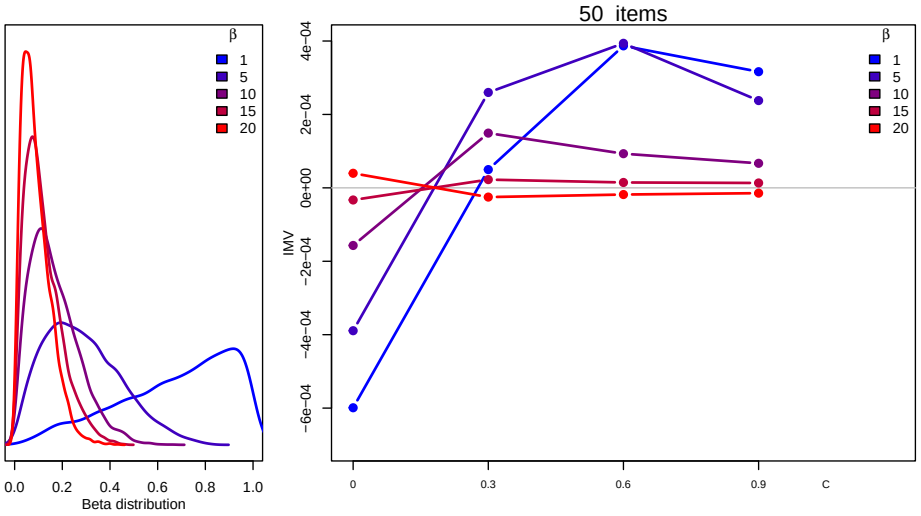
relative to the 1PL. The IMV(2PL,3PL) is positive but small and not increasing as a function of sample size.

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).

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.



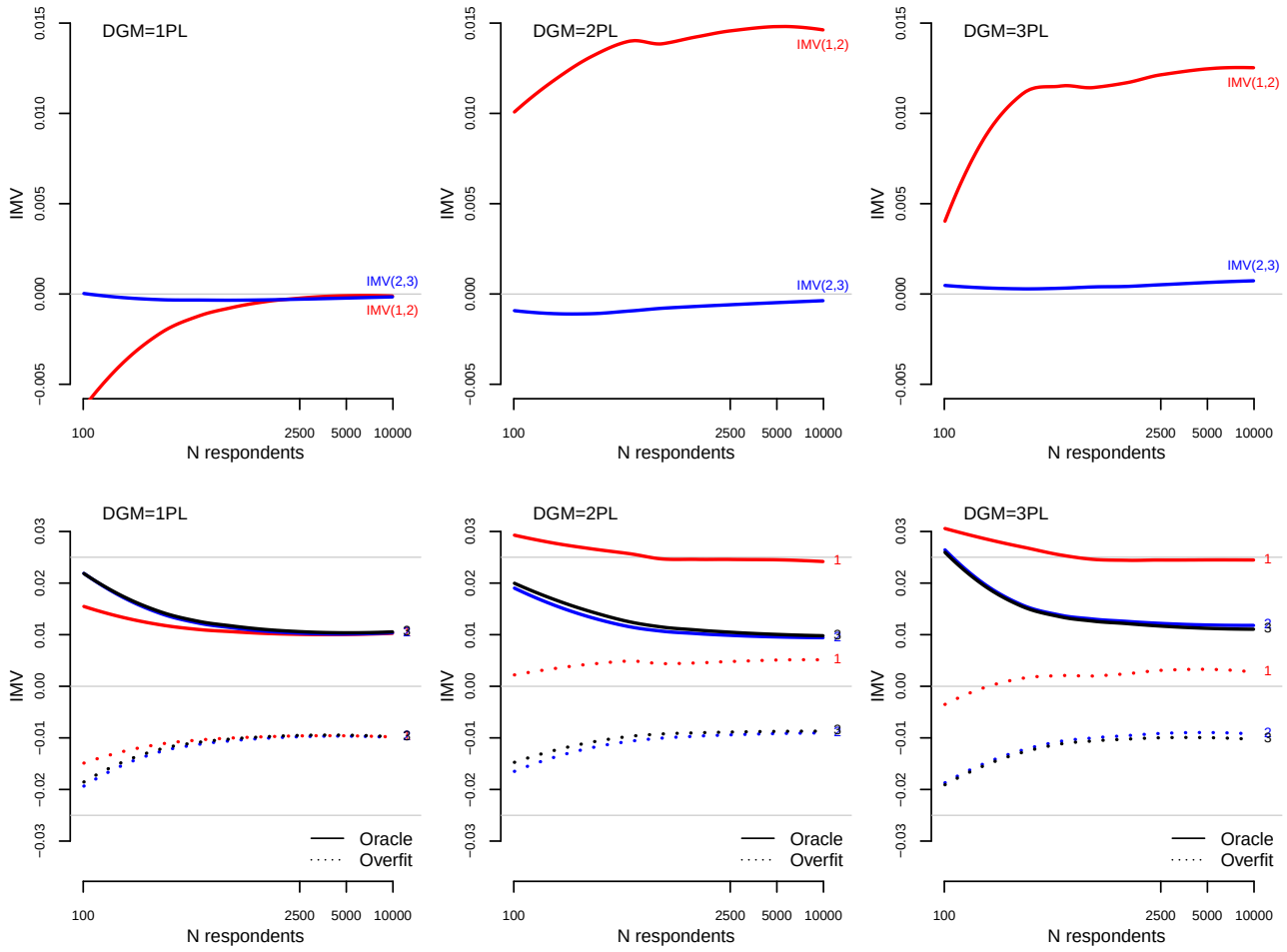
In Figure S8, the IMV does not go to zero for increasingly large samples but a fixed number of items. Why is this? The issue is that individual estimates of θ_i are largely unaffected by increases in the number of respondents thus leading to a floor in how accurate estimates of p_{ij} get for a given 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 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 et al., 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

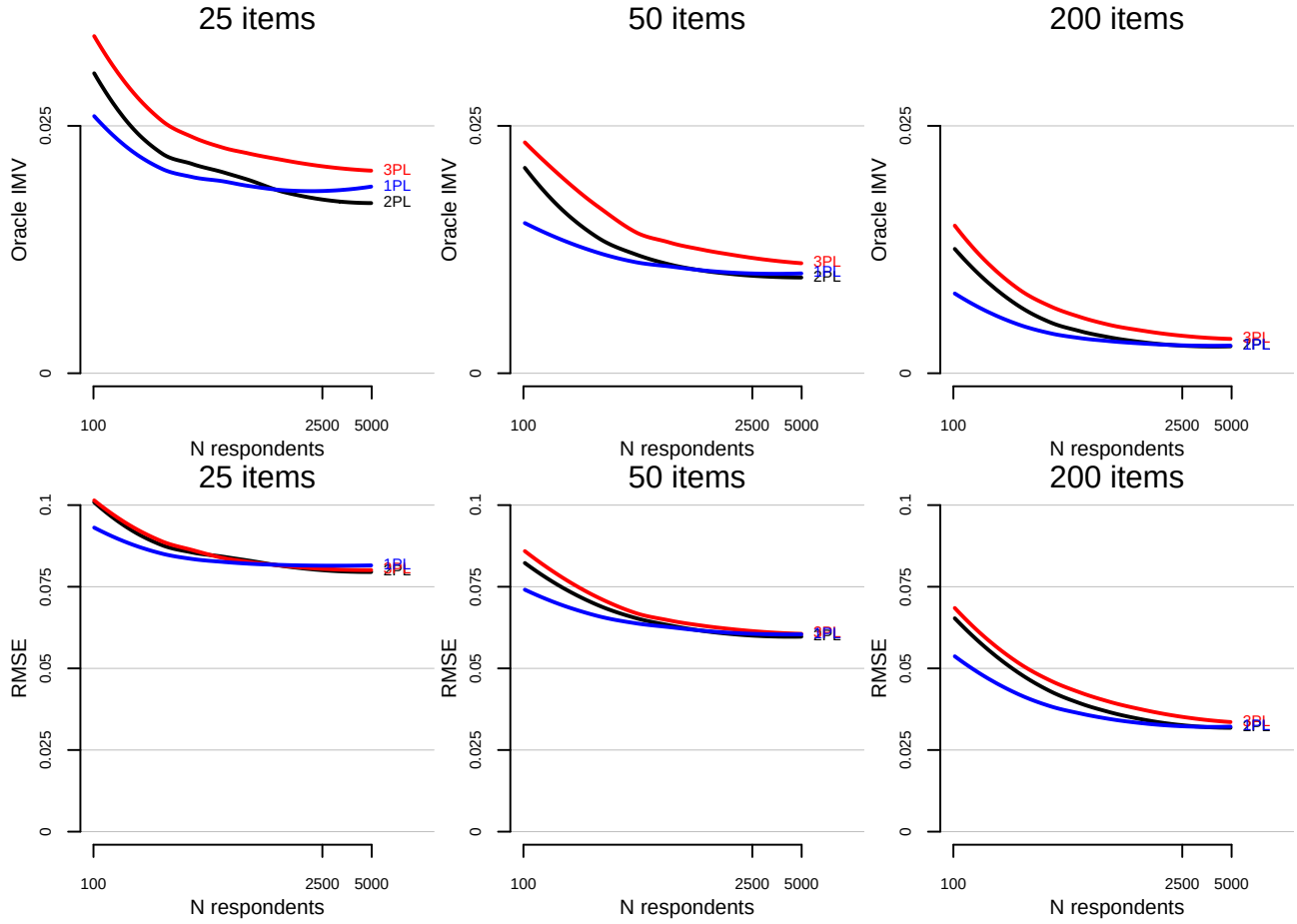
Figure S7. The IMV associated with different item response models (top) and the oracle/overfit values (bottom) for the 1PL/2PL/3PL for different choices of data generating model (DGM). We simulated data for 250 choices of $n \sim \text{Unif}(2, 4)$ where the number of respondents was 10^n with $C = 0.3$ 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.



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

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.



is apparent here, even absent the changes in sample size.

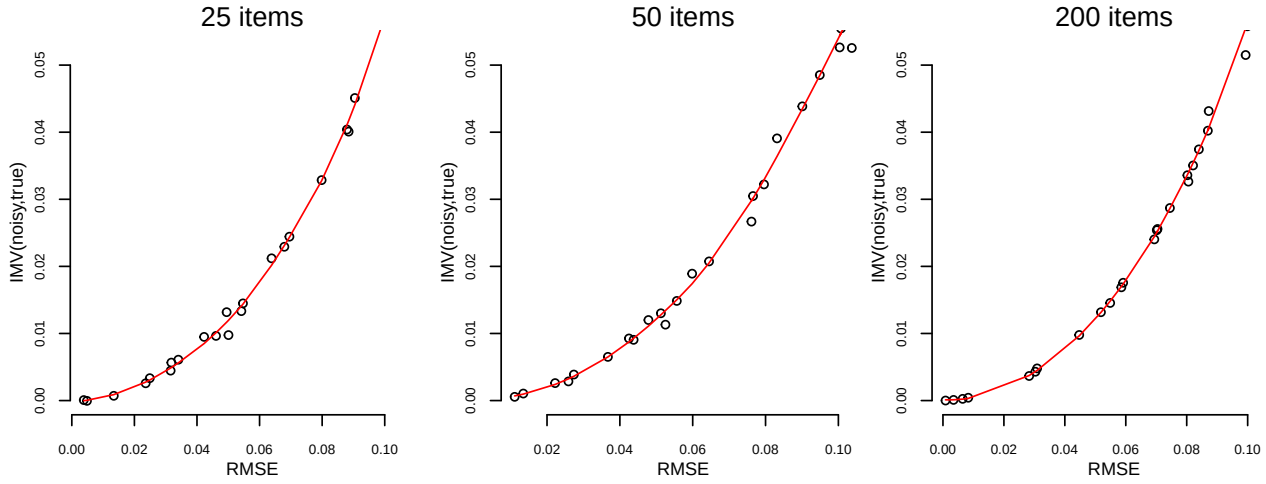
S2.8 Multidimensional models and the IMV

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 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

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 $\Delta \sim \text{Unif}(0,.2)$.



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 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.

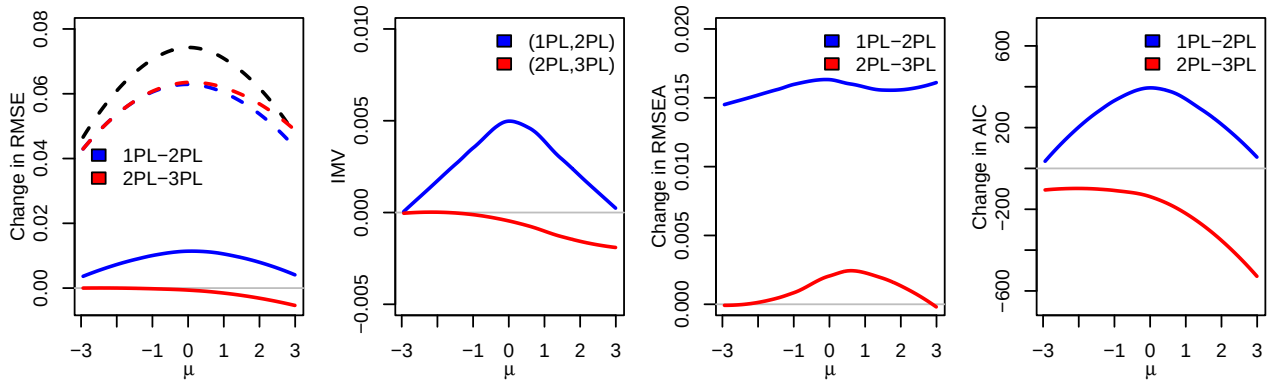
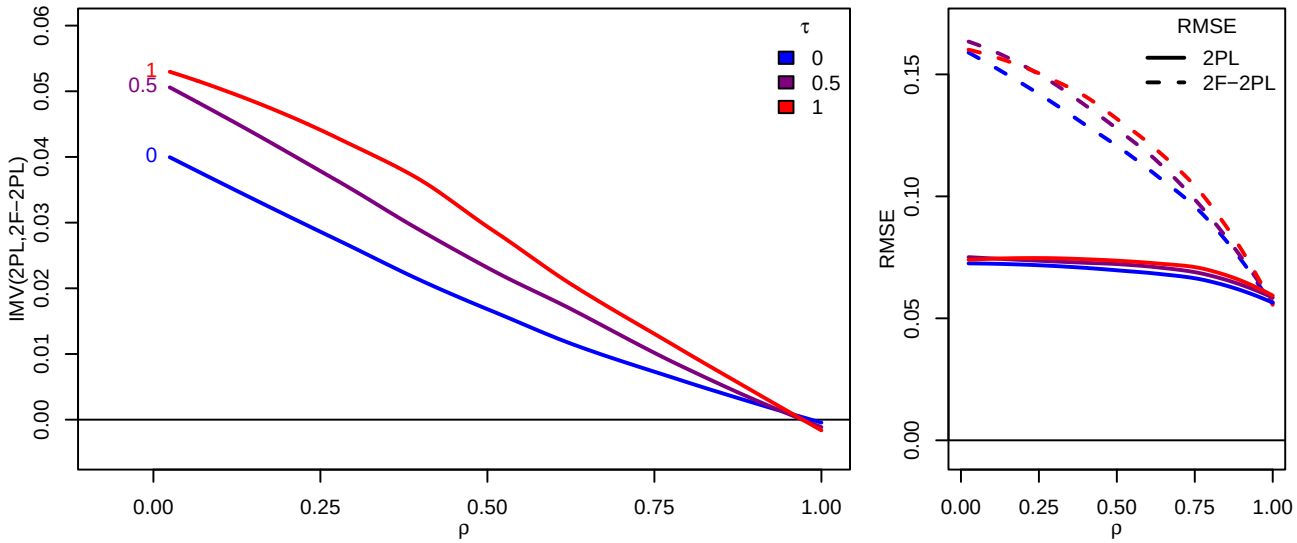


Figure S11. Results from multidimensional simulations. The left panel focuses on IMV(2PL,2F-2PL) as a function of ρ and τ. The right panels show the RMSE between the true and estimated probabilities separately for the 2PL and 2F-2PL models. Results are based on LOESS smoothing for 200 choices of ρ.



S3 Empirical Data

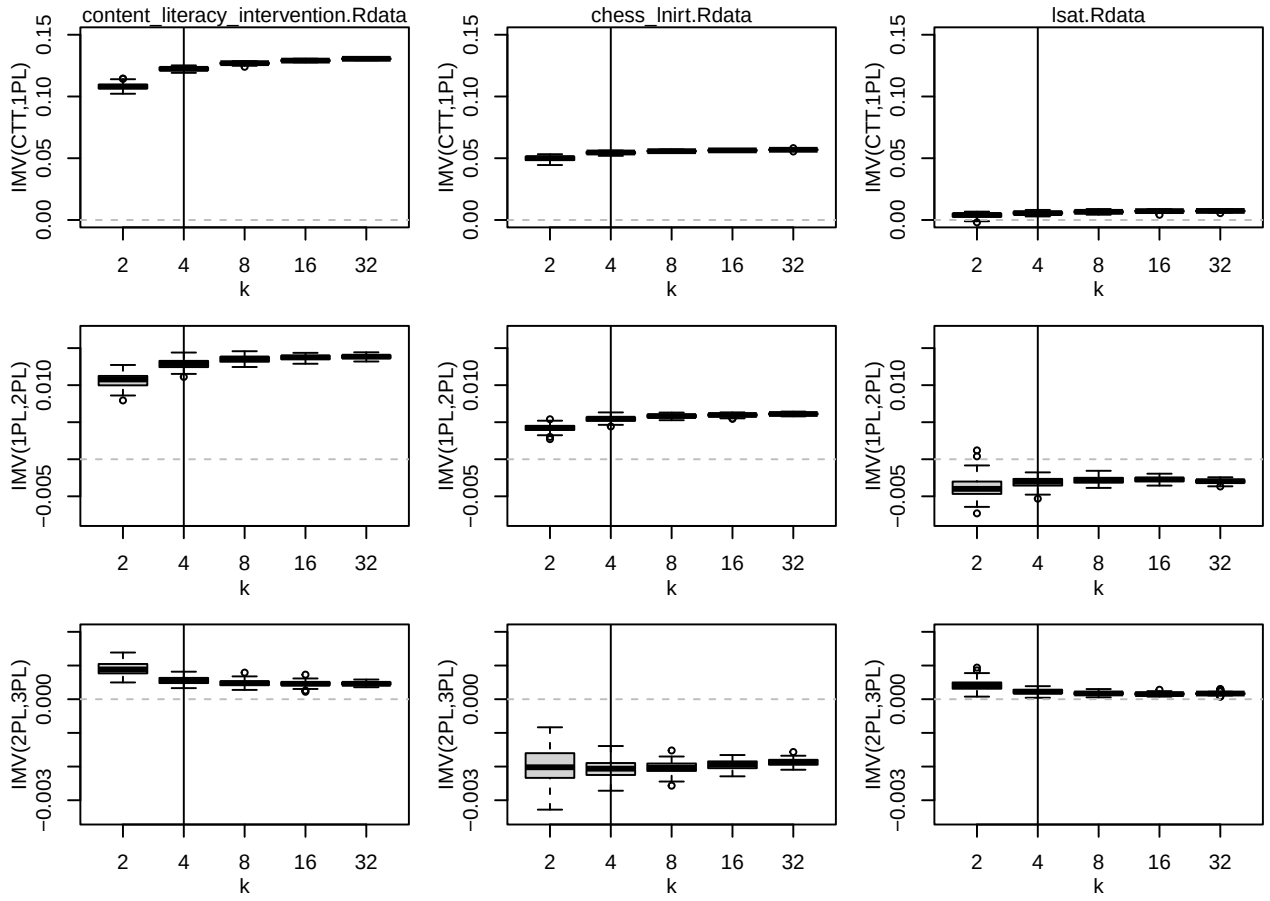
S3.1 Description of Data

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

Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, *60*(2), 255–265.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological methods & research*, *21*(2), 230–258.

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

Cai, L., Chung, S. W., & Lee, T. (2021). Incremental model fit assessment in the case of categorical data: Tucker–lewis index for item response theory modeling. *Prevention Science*, 1–12.

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

Craven, P., & Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische mathematik*, *31*(4), 377–403.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* ERIC.

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). Intermodel vigorish (imv): A novel approach for quantifying predictive accuracy when outcomes are binary. Retrieved from `https://osf.io/gu3ap/`

Doroudi, S. (2020). The bias-variance tradeoff: How data science can inform educational debates. *AERA Open*, *6*(4), 2332858420977208.

Eysenck, H. J., & Eysenck, S. B. (1968). Eysenck personality inventory. *Journal of Clinical Psychology*.

Feuerstahler, L. M. (2020). Metric stability in item response models. *Multivariate Behavioral Research*, 1–18.

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.

Guttman, L. (1950). The basis for scalogram analysis. *Measurement and prediction*, 60–90.

Haberman, S. J. (2005). Identifiability of parameters in item response models with unconstrained ability distributions. *ETS Research Report Series*, *2005*(2), i–22.

Haberman, S. J., Sinharay, S., & Lee, Y.-H. (2011). Statistical procedures to evaluate quality of scale anchoring. *ETS Research Report Series*, *2011*(1), i–20.

Han, Y., Zhang, J., Jiang, Z., & Shi, D. (2022). Is the area under curve appropriate for evaluating the fit of psychometric models? *Educational and Psychological Measurement*, 00131644221098182.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic

(roc) curve. *Radiology*, *143*(1), 29–36.

Haslbeck, J., & van Bork, R. (2024). Estimating the number of factors in exploratory factor analysis via out-of-sample prediction errors. *Psychological Methods*, *29*(1), 48–64.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., ... others (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kang, T., & Cohen, A. S. (2007). Irt model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331–358.

Köhler, C., Robitzsch, A., & Hartig, J. (2020). A bias-corrected rmsd item fit statistic: An evaluation and comparison to alternatives. *Journal of Educational and Behavioral Statistics*, *45*(3), 251–273.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. IAP.

Maris, G., & Bechger, T. (2009). On interpreting the model parameters for the three parameter logistic model. *Measurement*, *7*(2), 75–88.

Mavridis, D., Moustaki, I., & Knott, M. (2007). Goodness-of-fit measures for latent variable models for binary data. In *Handbook of latent variable and related models* (pp. 135–161). Elsevier.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 71–101.

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.

Maydeu-Olivares, A., & Garcia-Forero, C. (2010). Goodness-of-fit testing. *International encyclopedia of education*, *7*(1), 190–196.

Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2 n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*(471), 1009–1020.

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*.

Rahal, C., Verhagen, M., & Kirk, D. (2022). The rise of machine learning in the academic social sciences. *AI & Society*.

Reddy, S., Labutov, I., Banerjee, S., & Joachims, T. (2016). Unbounded human learning: Optimal scheduling for spaced repetition. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1815–1824).

Rizopoulos, D. (2006). ltm: An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. Retrieved from `https://doi.org/10.18637/jss.v017.i05`

Savalei, V., Brace, J., & Fouladi, R. T. (2021, May). *We need to change how we compute rmsea for nested model comparisons in structural equation modeling.* PsyArXiv. Retrieved from `psyarxiv.com/wprg8` doi: 10.31234/osf.io/wprg8

Savcisens, G., Eliassi-Rad, T., Hansen, L. K., Mortensen, L. H., Lilleholt, L., Rogers, A., ... Lehmann, S. (2023). Using sequences of life-events to predict human lives. *Nature Computational Science*, 1–14.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.

Shmueli, G. (2010). To explain or to predict? *Statistical science*, *25*(3), 289–310.

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology*, *22*(6), 786–809.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, *64*(4), 583–639.

Stenhaug, B., & Domingue, B. (2022). Predictive fit metrics for item response models. *Applied Psychological Measurement*.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 44–47.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589–617.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2006). 21 assessing the fit of item response theory models. *Handbook of statistics*, *26*, 683–718.

Van der Linden, W. J. (2017a). *Handbook of item response theory: Volume 2: Statistical tools.* CRC Press.

Van der Linden, W. J. (2017b). *Handbook of item response theory: Volume 3: Applications.* CRC press.

Van Maanen, L., Been, P., & Sijtsma, K. (1989). Problem solving strategies and the linear logistic test model. In *Mathematical psychology in progress* (pp. 267–287). Springer.

Verhagen, M. D. (2022). A pragmatist's guide to using prediction in the social sciences. *Socius*, *8*, 23780231221081702.

von Davier, M. (2009). Is there need for the 3pl model? guess what? *Measurement: Interdisciplinary Research and Perspectives*, *27*.

Wagenmakers, E.-J., & Farrell, S. (2004). Aic model selection using akaike weights. *Psychonomic bulletin & review*, *11*(1), 192–196.

Wainer, H. (2016). Discussion of david thissen's bad questions: An essay involving item response theory. *Journal of Educational and Behavioral Statistics*, *41*(1), 100–103.

Watts, D. J. (2014). Common sense and sociological explanations. *American Journal of Sociology*, *120*(2), 313–351.

Watts, D. J. (2017). Should social science be more solution-oriented? *Nature Human Behaviour*, *1*(1), 1–5.

Watts, D. J., Beck, E. D., Bienenstock, E. J., Bowers, J., Frank, A., Grubesic, A., ... Salganik, M. (2018). Explanation, prediction, and causality: Three sides of the same coin?

Wolfram, T., Tropf, F. C., & Rahal, C. (2022, May). *Short essays written during childhood predict cognition and educational attainment close to or better than expert assessment.* SocArXiv. Retrieved from `osf.io/preprints/socarxiv/a8ht9` doi: 10.31235/osf.io/a8ht9

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th ed.). Cengage Learning.

Wu, M., & Adams, R. J. (2013). Properties of rasch residual fit statistics. *Journal of Applied Measurement*.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.