Predictive Fit Metrics for Item Response Models

Applied Psychological Measurement
2022, Vol. 46(2) 136–155
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01466216211066603
journals.sagepub.com/home/apm
SAGE

Benjamin A. Stenhaug[1], and Benjamin W. Domingue[1]

Abstract

The fit of an item response model is typically conceptualized as whether a given model could have generated the data. In this study, for an alternative view of fit, "predictive fit," based on the model's ability to predict new data is advocated. The authors define two prediction tasks: "missing responses prediction"—where the goal is to predict an in-sample person's response to an in-sample item—and "missing persons prediction"—where the goal is to predict an out-of-sample person's string of responses. Based on these prediction tasks, two predictive fit metrics are derived for item response models that assess how well an estimated item response model fits the data-generating model. These metrics are based on long-run out-of-sample predictive performance (i.e., if the data-generating model produced infinite amounts of data, what is the quality of a "model's predictions on average?"). Simulation studies are conducted to identify the prediction-maximizing model across a variety of conditions. For example, defining prediction in terms of missing responses, greater average person ability, and greater item discrimination are all associated with the 3PL model producing relatively worse predictions, and thus lead to greater minimum sample sizes for the 3PL model. In each simulation, the prediction-maximizing model to the model selected by Akaike's information criterion, Bayesian information criterion (BIC), and likelihood ratio tests are compared. It is found that performance of these methods depends on the prediction task of interest. In general, likelihood ratio tests often select overly flexible models, while BIC selects overly parsimonious models. The authors use Programme for International Student Assessment data to demonstrate how to use cross-validation to directly estimate the predictive fit metrics in practice. The implications for item response model selection in operational settings are discussed.

Keywords

item response theory, fit, prediction, model comparison, cross-validation

A focal point of psychological measurement is item response data generated when persons respond to items (e.g., multiple choice items in educational assessments). Item response models are

[1]The Graduate School of Education at Stanford University, Stanford, CA, USA

Corresponding Author:
Benjamin A. Stenhaug, The Graduate School of Education at Stanford University, 450 Serra Mall, Stanford, CA 94305, USA.
Email: benastenhaug@gmail.com

statistical models fit to such item response data. As with most statistical models, more and less flexible versions of item response models are available. Consider the common family of unidimensional models dichotomous item responses: The one parameter logistic (1PL), the two parameter (2PL) logistic, and the three parameter (3PL) logistic models. The 1PL model is the least flexible, with just a difficulty parameter for each item (Rasch, 1960). The 3PL model is the most flexible, with a difficulty, discrimination, and guessing parameter for each item (Birnbaum, 1968).

Suppose that data are generated by a 3PL model (i.e., the data-generating model, abbreviated "DGM") and that both a 2PL model and 3PL model are estimated using this data. What does it mean for one of these models to "fit" the data? In item response theory research literature, fit is often defined by whether the model could have produced the data (DiTrapani, 2019). In essence, a variety of methods aim to test the null hypothesis that the model generated the data. For example, the M2 statistic compares the expected (according to the model) to the observed (by counting the data) moments of a contingency table (Maydeu-Olivares & Joe, 2005). If there is a mismatch between these moments, then the null hypothesis that the model could have generated the data is rejected and it is concluded that the model has poor fit. This view of model fit leads to the counterintuitive result that models are less likely to fit larger datasets: As sample size increases so does the statistical power with which trivial discrepancies between the data and the model can be identified (Steiger, 1990). Similarly, posterior predictive checks use the model to simulate data, use discrepancy measures to compare that simulated data to the observed data, and then conclude whether the model could have produced the observed data based on those discrepancy measures (Sinharay et al., 2006).

Item response model simulation studies, which are commonly used to guide usage in empirical settings, often take a similar view of fit. Luecht and Ackerman (2018) summarize the majority of these simulation studies as following the script, wherein (1) a DGM model is chosen (e.g., the 3PL), (2) item and person parameters are specified and item response data are simulated, (3) a variety of models are estimated using the simulated data, and (4) those models are compared. Luecht and Ackerman (2018) point out the inevitable conclusion that the model with the same parameterization as the DGM best fits the data. Going a step further, they remark that "one might even conclude that that result is axiomatic, thus eliminating the need to ever again see this type of IRT simulation study published" (p. 66).

As an example of such a simulation study, consider Kang and Cohen (2007) who evaluated the effectiveness of a variety of item response model comparison methods such as Akaike's information criterion (AIC; Akaike (1974)) and Bayesian information criterion (BIC; Schwarz et al. (1978)). They simulated data via the 3PL DGM, and fit 1PL, 2PL, and 3PL models to the simulated data. Finally, and this is crucial, they evaluated a model comparison method's (e.g., BIC) performance according to its ability to choose the 3PL model as the best fitting model. Their implicit assumption was that, by definition, if a 3PL model generated the data, then a 3PL model must best fit the data. After all, that model produced the data. One of their conclusions was that BIC performed poorly for data generated from a 3PL model because BIC preferred a simpler (than the 3PL) model. Other research on model comparison methods has also assumed that the prediction-maximizing model shares the same parameterization as the DGM: For example, Svetina and Levy (2016) negatively judged NOHARM-based methods—including RMSR and ALR—for detecting the dimensionality of item response data, based on their tendency to find fewer than the data-generating number of dimensions at low sample sizes.

An Alternative Approach to Fit: Predictive Fit

The previous work to illustrate how fit is often conceptualized in the item response theory research literature is summarized. An alternative approach is predictive fit which is based on how well a model predicts new (i.e., out-of-sample) data from the DGM (Gelman et al., 2014). The fundamental logic of predictive fit is that the model with the best predictions is likely to be the most useful. When introducing the now eponymous model, Rasch (1960) wrote, “When you construct a model you leave out all the details Models should not be true, but it is important that they are” (p. 38). Lord (1983) argued that the Rasch model should be preferred at small sample sizes, even if it is known to be the “wrong” model, precisely because it might offer better predictions. At the core of these suggestions is that, the goal of model selection should be to identify an effective model, not necessarily the model that generated the data.

One might argue that item response models are typically used to explain and understand as opposed to predict. The authors agree with the many who have argued that this is a false dichotomy and that explanation and prediction are in fact two sides of the same coin (Watts et al., 2018; Yarkoni & Westfall, 2017). Feynman et al. (1965) argued that science is the process of making hypotheses and checking their predictions against new data. In other words, the ultimate test of a scientific hypothesis is its ability to make predictions. Focusing on item response data, the conclusions from item response models that make poor predictions should not be trusted. In fact, in operational settings, one cannot know that the data came from an item response model; the question is rather whether item response models can characterize the data usefully. The better an item response model’s predictions, the better it has characterized the data and the more trust we can place in its conclusions. Further, we argue that many item response model simulation studies would be more valuable if they assessed models according to their predictive fit. The predictive fit view argues that it is better to have a model that produces high-quality predictions than it is to have a model with the same parameterization as the DGM. Thus, Kang and Cohen (2007) might have judged model selection methods not by their ability to identify the DGM, but instead by their ability to select the model that makes the best predictions.

In essence, a compelling way is argued to assess how applicable or useful an item response model is by the quality of its predictions. Accordingly, our goal is to forward the predictive fit view by taking a step back and delineating two distinct prediction tasks for an item response model. The first prediction task, which is named name “missing responses,” is to predict the probability of a missing item response. The second prediction task, which is named “missing persons,” is to predict the probability of all of the responses from a new, randomly drawn person. These two prediction tasks correspond to two predictive fit metrics, which is defined as measures of how well an item response model predicts new data from the DGM.

Organization. The background on item response models and how they are typically compared in practice is given first. Second, the two possible prediction tasks which correspond to different definitions of out-of-sample for item response data are described. Third, two predictive fit metrics based on these two definitions are derived. These metrics are derived for the theoretical case when the DGM is known, such as in a simulation study. Fourth, the behavior and utility of these metrics are shown in four simulation studies. To reexamine and extend Kang and Cohen (2007), the prediction-maximizing model to the model selected by common methods such as AIC and BIC is compared. Lastly, a real world example of model selection is provided, which includes a description of how to use cross-validation to estimate the predictive fit metrics in practice.

Item Response Models

Let Y represent an observed item response matrix. y_{ij} is an observed dichotomous item response where $y_{ij} = 1$ indicates that the ith person responded correctly to the jth item and $y_{ij} = 0$ indicates that they responded incorrectly. Item response theory provides a framework for modeling Y, especially when there is a reason to believe that one or more latent traits exists. The fundamental building block of item response theory is the item response function (IRF) which gives the probability that a person will respond correctly to (or positively endorse) an item (Baker & Kim, 2004). The 3PL IRF is commonly used and is specified as

$$Pr\left(y_{ij} = 1\right) = c_j + \left(1 - c_j\right)F\left(a_j\theta_i + b_j\right)$$

where θ_i is the ith person's ability; a_j, b_j, and c_j are the jth item's discrimination, easiness, and guessing parameters respectively; and F is the sigmoid function, $F(x) = e^x/1 + e^x$. The two parameter logistic (2PL) and one parameter logistic (1PL) IRFs can be thought of as constrained forms of the 3PL IRF. The 2PL IRF constrains the guessing parameter c_j to 0. The 1PL IRF constrains the guessing parameter c_j to 0 and the discrimination parameter a_j to 1.

Item response models are commonly estimated using marginal maximum likelihood estimation (MMLE), which estimates item parameters while treating person abilities as a nuisance parameter (Bock, 1983; Casabianca & Lewis, 2015). MMLE was developed to remedy the fact that simultaneously estimating item and person parameters yields statistically inconsistent estimates as the number of persons goes to infinity (Bock & Aitkin, 1981). As such, MMLE estimates item parameters by maximizing the marginal log likelihood (MLL) of the model

$$MLL(model(Y)) = \sum_{i}^{I} log \int \left[\prod_{j=1}^{J} \widehat{Pr}\left(y_{ij}\middle|\widehat{\boldsymbol{\psi}}_{\mathbf{j}},\boldsymbol{\theta}\right)\right]\widehat{g}(\boldsymbol{\theta})d\boldsymbol{\theta}$$

where $\widehat{\boldsymbol{\psi}}_{\mathbf{j}}$ is the vector of estimated item parameters and $\widehat{g}(\boldsymbol{\theta})$ is the distribution of ability. Typically, $\widehat{g}(\boldsymbol{\theta})$ is assumed to be normally distributed with hyperparameters (e.g., mean and/or variance) estimated during model fitting. (Baker & Kim, 2004). If necessary, person ability estimates can be obtained using an estimation technique such as expected a-posteriori (EAP) or maximum a-posteriori (MAP) following MMLE of item parameters (Bock, 1983).

Item Response Model Selection In Practice

Item response models are typically compared using likelihood ratio tests (LRT) or information criterion such as AIC and BIC (Maydeu-Olivares, 2013). Each of these methods is based on the model's marginalized log likelihood, $MLL(model(Y))$, and number of item parameters estimated in fitting the model, M. Consider two models fit to Y: $model_1(Y)$ and $model_2(Y)$ where the latter model has a greater number of parameters such that $M_2 > M_1$. An LRT compares these models by exploiting the fact that

$$2 \cdot [MLL(model_2(Y)) - MLL(model_1(Y))]$$

follows a chi-squared distribution with $M_2 - M_1$ degrees of freedom (Andersen, 1973; Baker & Kim, 2004). If the associated p-value is statistically significant (often at a 0.05 significance level), then it is concluded that $model_2(Y)$ fits better than $model_1(Y)$. On the other hand, AIC and BIC both add a penalty to the likelihood based on M

$$AIC(model) = -2 \cdot MLL(model(Y)) + 2M$$
$$BIC(model) = -2 \cdot MLL(model(Y)) + log(I) \cdot M$$

where I is the number of persons (i.e., rows) in Y. Lower values of AIC and BIC indicate better fit. With the goal of defining predictive fit metrics, the missing responses and missing persons prediction tasks for item response models are now described from first principles.

Out-of-Sample for Item Response Data

The goal of predictive fit metrics is to measure how well a model predicts out-of-sample data from the DGM, but what, exactly, should be considered out-of-sample? Should it be the person, the item, or the item response that is out-of-sample? The fact that item responses are cross-classified within persons and items complicates this discussion (Furr, 2017). If entire persons are out-of-sample, then in-sample ability estimates are unavailable, meaning that they cannot be used to generate predictions. Instead, predictions can be made based on $\widehat{g}(\boldsymbol{\theta})$. On the other hand, if it is single item responses that are out-of-sample, then a person's responses to in-sample items can be used to generate in-sample ability estimates.

We denote some arbitrary out-of-sample matrix $\tilde{Y}$. Two[1] versions of $\tilde{Y}$ are considered, which vary based on what is considered out-of-sample. The first version of $\tilde{Y}$ comes from in-sample persons responding to in-sample items. This out-of-sample matrix as $\tilde{Y}^{MR}$ is denoted, with "MR" abbreviating "missing responses." The unit of observation for $\tilde{Y}^{MR}$ is the item response. The missing response on the left of Figure 1 shows that Person A's response to item 1 is missing. The model's prediction task is to estimate the probability of this missing response. To do so, the model can use the other persons to estimate item 1's parameters and the other items to estimate Person A's ability. This logic can be applied to each entry $\tilde{Y}^{MR}$, and therefore $\tilde{Y}^{MR}$ has the same dimensions as Y. Adaptive testing is an application in which the missing responses prediction task might make sense: The goal of an adaptive testing engine is often to next assign an item that the person has a fixed chance (e.g., 50%) of responding correctly to. Accordingly, the model that can best estimate these probabilities may be most useful.

The second version of $\tilde{Y}$ comes from out-of-sample persons responding to in-sample items. This out-of-sample matrix as $\tilde{Y}^{MP}$ is denoted, with "MP" abbreviating "missing persons." The unit of observation for $\tilde{Y}^{MP}$ is a person's vector of item responses. The bottom row on the right of Figure 1 represents a new person, Person D, responding to each of the items for the first time. The prediction task is for the model to estimate the likelihood of all of Person D's item responses. The other persons is used to estimate item parameters, but there is no way to estimate Person D's ability. As a result, a prediction about their entire vector of item responses has to be made—the unit of analysis—by treating ability as a nuisance variable; to do this, we average (i.e., integrate) over the distribution, denoted $g(\theta)$, from which we assume Person D's ability originates. So that $\tilde{Y}^{MP}$ has the same scale as Y, the authors might consider there to be as many missing persons as there are persons in Y. Traditional linear testing is an application wherein the missing persons prediction task might be preferred: the ability of the next test taker is unknown and a reasonable goal might be to prefer a scoring model that will perform optimally in the case of this heightened uncertainty. Note that LRT, AIC, and BIC all implicitly target the missing persons prediction task because they are based on the marginalized log likelihood of the model.

Predictive Fit Metrics

A predictive fit metric for each of $\tilde{Y}^{MR}$ and $\tilde{Y}^{MP}$ is now derived. These metrics can only be calculated exactly when the DGM is known such as in a simulation study (how to use cross-validation to estimate these metrics in practice is described later). When the DGM is known, a model's predictive performance on the produced can be directly measured by the DGM. Conceptually, this is equivalent to using the DGM to simulate an infinite amount of out-of-sample data, and then measuring a model's fit to the DGM based on its predictive performance for this (infinite) out-of-sample data. In particular, the metrics measure how well a model predicts all possible out-of-sample matrices that the DGM might produce, weighted by their probability of being produced. Both metrics begin with the likelihood of a single $\tilde{Y}$ according to a model fit to Y, which is generically denoted as $model(Y)$. This is known as log predictive likelihood (lpl), which can be thought of as a function that takes $\tilde{Y}$ and a model fit to Y as inputs and outputs the log likelihood of $\tilde{Y}$ according to that model (Gelman et al., 2014)

$$lpl\left(\tilde{Y}, model(Y)\right) = log\widehat{Pr}\left(\tilde{Y}\middle|model(Y)\right).$$

Metric 1: Expected Log Predictive Likelihood for Missing Responses (ELPL-MR)

Calculation of log predictive likelihood of missing responses (lpl-MR) for a single $\tilde{Y}^{MR}$ is relatively straightforward because we can use estimates of person abilities so that

$$lpl - MR\left(\tilde{Y}^{MR}, model(Y)\right) = log\widehat{Pr}\left(\tilde{Y}^{MR}\middle|model(Y)\right) = \sum_{i=1}^{I}\sum_{j=1}^{J} log\widehat{Pr}\left(\tilde{y}_{ij}^{MR}\middle|\widehat{\boldsymbol{\psi}}_{\mathbf{j}},\widehat{\boldsymbol{\theta}}_{\mathbf{i}}\right)$$

where $\tilde{y}_{ij}^{MR}$ is an item response from Y^{MR}, $\widehat{\boldsymbol{\psi}}_{\mathbf{j}}$ is item j's vector of parameter estimates from $model(Y)$, and $\widehat{\boldsymbol{\theta}}_{\mathbf{i}}$ is person i's vector of ability estimates from $model(Y)$. We use the short-hand $\widehat{p}_{i,j}^{MR} = \widehat{Pr}(y_{i,j}^{MR} = 1|\widehat{\boldsymbol{\psi}}_{\mathbf{j}},\widehat{\boldsymbol{\theta}}_{\mathbf{i}})$. To be concrete, in the case of the unidimensional 2PL model specification

$$lpl - MR\left(\tilde{Y}^{MR}, model(Y)\right) =$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\tilde{y}_{ij}^{MR}log\left(\widehat{p}_{i,j}^{MR}\right) + \left(1-\tilde{y}_{ij}^{MR}\right)log\left(1-\widehat{p}_{i,j}^{MR}\right) =$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\tilde{y}_{ij}^{MR}log\left(F\left(\widehat{a}_j\widehat{\theta}_i + \widehat{b}_j\right)\right) + \left(1-\tilde{y}_{ij}^{MR}\right)log\left(1-F\left(\widehat{a}_j\widehat{\theta}_i + \widehat{b}_j\right)\right).$$

Of course, there are many possible out-of-sample item response matrices $\tilde{Y}^{MR}$. The measure of model performance should be reflective of the DGM in general, not one particular $\tilde{Y}^{MR}$. The DGM is captured simply by $p_{i,j}^{MR}$, the data-generating probability of y_{ij}^{MR} being responded to correctly. If the DGM is an item response model, $p_{i,j}^{MR} = Pr(y_{i,j}^{MR} = 1|\boldsymbol{\psi}_{\mathbf{j}},\boldsymbol{\theta}_{\mathbf{i}})$. The out-of-sample predictive performance metric of interest is Expected Log Predictive Likelihood for Missing Responses (ELPL-MR), which is the expectation of lpl

$$\begin{aligned} ELPL - MR(model(Y)) &= \mathbb{E}\left[lpl\left(\tilde{Y}^{MR}, model(Y)\right)\right] \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}p_{i,j}log\left(\widehat{p}_{i,j}\right) + (1-p_{i,j})log\left(1-\widehat{p}_{i,j}\right). \end{aligned}$$

In essence, ELPL-MR can be thought of as a function that takes a model fit to Y as input and outputs the expectation of the log likelihood of $\tilde{Y}^{MR}$.

One way to think about this equation is that ELPL-MR is the weighted average of the log likelihood, where the weights are determined by the true probabilities. As a simple example, consider a model that predicts that an item response will be correct at a rate of 0.8 but the true data-generating probability is 0.9. The long-run log likelihood of the item response according to the model is $0.9log0.8 + 0.1log0.2 \approx -0.36$. Translating back to the probability scale, the long-run likelihood is $exp(-0.36) \approx 0.70$.

Metric 2: Expected Log Predictive Likelihood for Missing Persons (ELPL-MP)

The predictive fit metric is now derived for when the prediction task is the vector of responses for persons not known to the model as is a row vector, $\mathbf{y}_{\mathbf{u}}^{MP}$, from $\tilde{Y}^{MP}$. Calculation of log predictive likelihood of missing persons (lpl-MP) for $\tilde{Y}^{MP}$ is complicated by the fact that the persons in $\tilde{Y}^{MP}$ are out-of-sample and therefore unobserved in Y; hence, ability estimates are unavailable. However, as is standard in MMLE, we can calculate a marginalized likelihood by taking the expectation over $\widehat{g}(\boldsymbol{\theta})$, the distribution of ability as estimated by the model. We begin by calculating the lpl of $\mathbf{y}_{\mathbf{u}}^{MP}$

$$lpl - MP\left(\mathbf{y}_{\mathbf{u}}^{MP},model(Y)\right) = \int \widehat{Pr}\left(\mathbf{y}_{\mathbf{u}}^{MP}|\boldsymbol{\theta}\right)\widehat{g}(\boldsymbol{\theta})d\boldsymbol{\theta} = \int \left[\prod_{j=1}^{J} \widehat{Pr}\left(y_{uj}^{MP}\right)\middle|\widehat{\boldsymbol{\psi}}_{\mathbf{j}},\boldsymbol{\theta}\right)\right]\widehat{g}(\boldsymbol{\theta})d\boldsymbol{\theta}$$

Next, the data-generating distribution of $\tilde{Y}^{MP}$ is to be accounted for, which is captured by π_u, the probability of a random person from the DGM producing y_u. There are U possible response patterns (e.g., a dichotomous test with J items has $U = 2^J$ possible response patterns). Assuming the DGM is an item response model, we calculate π_u as follows

$$\pi_u = \int Pr\left(\mathbf{y}_{\mathbf{u}}^{MP}|\boldsymbol{\theta}\right)g(\boldsymbol{\theta})d\boldsymbol{\theta} = \int \left[\prod_{j=1}^{J} Pr\left(y_{uj}^{MP}\right)\middle|\boldsymbol{\psi}_{\mathbf{j}},\theta\right)\right]g(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

The out-of-sample predictive performance metric of interest is Expected Log Predictive Likelihood for Missing Persons (ELPL-MP), which is the expectation of lpl over each possible $\mathbf{y}_{\mathbf{u}}$

$$\begin{aligned} ELPL - MP(model(Y)) &= \mathbb{E}\left[lpl\left(\mathbf{y}_{\mathbf{u}}^{MP},model(Y)\right)\right] \\ &= \sum_{u=1}^{U} \pi_u \cdot lpl\left(\mathbf{y}_{\mathbf{u}}^{MP},model(Y)\right) \\ &= \sum_{u=1}^{U} \pi_u \cdot \left(\int \left[\prod_{j=1}^{J} \widehat{Pr}\left(y_{uj}\middle|\widehat{\boldsymbol{\psi}}_{\mathbf{j}},\theta\right)\right]\widehat{g}(\theta)d\boldsymbol{\theta}\right) \end{aligned}$$

As with ELPL-MR, ELPL-MP can be thought of as a function that takes a model fit to Y as input and outputs the expectation of the log likelihood of $\tilde{Y}^{MP}$. In practice, integrals can be approximated using Gauss–Hermite quadrature (Embretson & Reise, 2013).

Simulation Studies

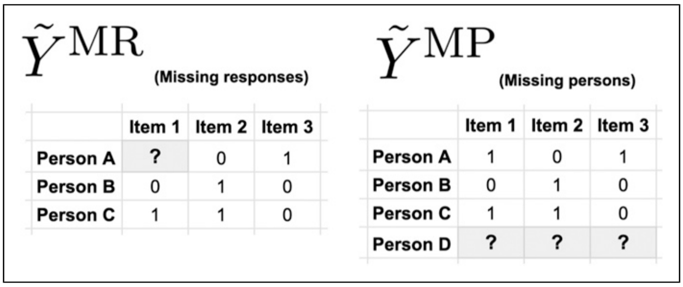
To demonstrate the behavior and utility of the two predictive fit metrics, ELPL-MR and ELPL-MP, four simulation studies were conducted. The first revisited Kang and Cohen (2007) using predictive fit. The second and third both used a 3PL DGM and explored the role different ability distributions, sample sizes, and item architectures play in which of the 1PL, 2PL, and 3PL model have the best predictive fit. The first three simulation studies used exclusively unidimensional (a single ability factor); the fourth compared models with varying numbers of factors.

Throughout, the prediction-maximizing model was identified according to the two predictive fit metrics. Calculations were made on which model AIC, BIC, and LRT should be selected in order to demonstrate how they perform vis-a-vis our two predictive metrics. Each simulation study involved many replications where a replication consisted of the following steps: (1) simulate data using the DGM, (2) fit a variety of models to the simulated data, (3) calculate ELPL-MR, ELPL-MP, the likelihood, AIC, and BIC for each of the models, and (4) determine the winning model according to each of these methods. For ELPL-MR and ELPL-MP, the model with the greatest value is the prediction-maximizing model. For AIC and BIC, the model with the lowest value is the selected model. For LRT, we conducted a sequence of tests so that, for example, if the 2PL model was statistically significant compared to the 1PL model but not the 3PL model, then the 2PL model was selected. In general, let us say a model "wins" if it is optimal according to a specific measure. The DGM and models fit in steps (1) and (2) varied across simulation studies; steps (3) and (4) were consistent across the simulation studies.

R was used for computing (R Core Team, 2019). The R package, mirt, was used to fit models using MMLE with the EM algorithm and 61 quadrature points (Chalmers, 2012) (Figure 1). We used custom written functions to calculate ELPL-MR and ELPL-MP for each model. In particular, ELPL-MR was calculated using equation. Abilities were estimated using both MAP and EAP with the usual standard normal prior. Because results using EAP and MAP were nearly identical, only results using EAP ability estimates are reported.[2] ELPL-MP was calculated using equation. Integrals were approximated using Gauss-Hermite quadrature with 61 points (Embretson & Reise, 2013). We used the suite of R packages known as the tidyverse for data wrangling and visualization (Wickham, 2017). Materials to reproduce this article, including functions to estimate ELPL-MR and ELPL-MP, are available at github.com/stenhaug/irt-predictive-fit.

Methods for Simulation Study 1

In Simulation Study 1, Kang and Cohen (2007) were revisited who evaluated model selection methods (e.g., BIC) via their capacity to identify the model with the same parameterization as the DGM (e.g., a model selection method should choose the 3PL model if the 3PL DGM was used) (Table 1). Whether the 3PL model actually had the best predictive fit in the conditions in which they conducted their simulation study was questioned. The six conditions from Kang and Cohen (2007) that came from crossing the DGM (1PL, 2PL, or 3PL) and sample size (500 or 1000 persons) were focused. In each condition, 20 items were used and drew abilities from a normal distribution, $\theta \sim N(0,1)$. Their exact item parameters are used as reported in Table 4 of Kang and Cohen (2007). Five hundred replications for each condition were conducted. Our hypothesis was that the winning model—that is, the most predictive model—would not always have the same parameterization as the DGM. The authors believed that this was most likely at lower sample sizes where more complex models might overfit the data, leading to poor predictions.



$\tilde{Y}^{\text{MR}}$ (Missing responses)

	Item 1	Item 2	Item 3
Person A	?	0	1
Person B	0	1	0
Person C	1	1	0

$\tilde{Y}^{\text{MP}}$ (Missing persons)

	Item 1	Item 2	Item 3
Person A	1	0	1
Person B	0	1	0
Person C	1	1	0
Person D	?	?	?

Figure 1. Understanding the two out-of-sample item response matrices, $\tilde{Y}^{MR}$ and $\tilde{Y}^{MP}$. With missing responses, the unit of observation is a single item response and the person's other responses can be used to estimate ability. With missing persons, the unit of observation is the person's response vector and there are no responses with which to estimate ability.

Table 1. Simulation Study 1 results.

		Estimated Model														
		ELPL-MR			ELPL-MP			BIC			AIC			LRT		
DGM	Persons	1PL	2PL	3PL	1PL	2PL	3PL	1PL	2PL	3PL	1PL	2PL	3PL	1PL	2PL	3PL
1PL	500	500	0	0	500	0	0	500	0	0	497	3	0	480	20	0
1PL	1000	500	0	0	500	0	0	500	0	0	499	1	0	482	182	0
2PL	500	0	499	1	0	499	1	0	500	0	0	0	500	0	500	0
2PL	1000	0	498	2	0	498	0	0	500	0	0	0	500	0	499	0
3PL	500	0	488	12	0	245	255	86	414	0	0	388	112	0	246	254
3PL	1000	0	407	93	0	11	489	0	500	0	0	136	364	0	44	456

Results for Simulation Study 1

Table 1 shows the number of replications in which each model won. In the conditions with the 1PL or 2PL DGM, the prediction-maximizing model according to ELPL-MR and ELPL-MP nearly always shared the DGM's parameterization. In these conditions, the model selection methods, BIC, AIC, and LRT, nearly always selected this model with the exception that the LRT occasionally (4% of replications) chose the 2PL model when the prediction-maximizing model was the 1PL model. In contrast, with the 3PL DGM, the 2PL model often won based on our predictive metrics (i.e., the 2PL optimally predicted the out-of-sample data). Under these conditions, Kang and Cohen (2007) found that AIC often selected the 2PL model, which they interpreted as a failure of AIC. Our results instead indicate that if the goal is to identify the model with the greatest predictive fit, AIC may actually have been a useful metric. In general, LRT selects models consistent with ELPL-MP, while AIC and BIC selects models more consistent with ELPL-MR.

The model with the greatest ELPL-MR was often simpler than the model with the greatest ELPL-MP. For example, with a 3PL DGM and 500 persons, the 2PL model outperformed the 3PL model in 488 out of 500 replications according to ELPL-MR and in 245 out of 500 replications according to ELPL-MP. This is taken as evidence that ELPL-MR prefers more parsimonious models than ELPL-MP. Why is this so? Recall that the difference between ELPL-MP and ELPL-MR is how they treat ability. ELPL-MP assumes ability to be coming from a generic distribution, $g(\boldsymbol{\theta})$, whereas ELPL-MR actually estimates each person's ability. As a result, ELPL-MR requires estimation of more parameters (item parameters and a parameter for each person) than ELPL-MP (just item parameters). Estimation of additional parameters requires increased sample size. When ELPL-MR is calculated, the additional step of estimating each person's ability is taken, which causes the imperfection in the item parameter estimates to propagate to the person abilities. On the other hand, when ELPL-MP is calculated, $g(\boldsymbol{\theta})$ is just integrated over which is much more tolerant of those imperfect item parameter estimates.[3]

Methods for Simulation Study 2

Simulation Study 1 found that with the 3PL DGM, the 2PL model was frequently best according to predictive performance metrics, especially if the number of persons was relatively small (Figure 2). Simulation Study 2 builds on this observation by exploring the role of sample size (i.e., number of persons) and ability distribution in determining which model best fits a 3PL DGM. In Simulation Study 2, we used the 3PL DGM, 20 items, and item parameters from Kang and Cohen (2007). Two thousand replications were conducted, each of which was as follows. The number of persons were drawn from a discrete uniform distribution, $I \sim unif\{100, 10000\}$. Abilities were

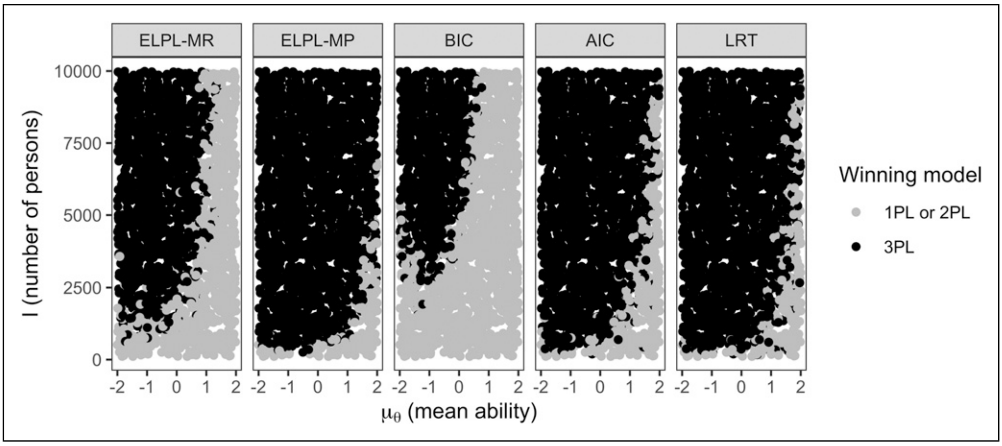


Figure 2. Simulation study 2 results. Each point corresponds to the prediction-maximizing model according to the predictive fit metrics, ELPL-MR and ELPL-MP, or the model selected by BIC, AIC, or LRT for one of 2000 replications. The 3PL model was most likely to offer the best fit and be selected with more persons and at lower mean ability (guessing is more prominent). For the predictive fit metrics, ELPL-MP preferred more flexible models than ELPL-MR. For the model selection methods, LRT and AIC preferred more flexible models than BIC.

drawn from a normal distribution, $\theta_i \sim N(\mu_\theta,1)$, where the mean of that distribution was drawn from a continuous uniform distribution, $\mu_\theta \sim unif(-2, 2)$. As before, data were simulated using these parameters, fit the 1PL, 2PL, and 3PL models, determined the best fitting model according to ELPL-MR and ELPL-MP, and identified the model selected by BIC, AIC, and LRT.

Results for Simulation Study 2

Figure 2 shows the prediction-maximizing model according to ELPL-MP and ELPL-MR as well as the model selected by BIC, AIC, and LRT. As in Simulation Study 1, ELPL-MR preferred more parsimonious models as evidenced by the 2PL model winning more frequently according to ELPL-MR than according to ELPL-MP. As anticipated, the greater the number of persons, I, the more likely the 3PL model was to win. However, the ability distribution is also salient. As μ_θ increased, the 3PL became less likely to win. This is to be expected; guessing plays less of a role for high ability persons, which decreases the predictive value of the model including a guessing parameter. LRT and AIC tended to select the prediction-maximizing model according to ELPL-MP. BIC selected models more in line with the prediction-maximizing model according to ELPL-MR.

Figure 2 can be read in terms of minimum sample requirements for the 3PL model (although results may depend somewhat on the specific set of item parameters). When μ_θ is less than 0, the sample size at which the 3PL model tended to outperform the 2PL model was somewhat low (≈2000) according to ELPL-MP, and it was a bit higher according to ELPL-MR. As μ_θ increased, the relative predictive performance of the 3PL model decreased quickly, so much so that, for ELPL-MR, the 3PL model nearly never won when μ_θ was greater than one.

Methods for Simulation Study 3

Simulation Studies 1 and 2 both used item parameters from Kang and Cohen (2007). In Simulation Study 3, item parameters were simulated with the goal of understanding how different item architectures—the distribution of easiness, discrimination, and guessing—effect which model wins according to ELPL-MR and ELPL-MP (Figure 3). Simulation Study three again exclusively used the 3PL DGM. Nine conditions were first created by crossing the vector of guessing parameters c (set to 0.03, 0.10, 0.25 for all items) and the sample size (set to 1000, 5000, or 10,000 persons). One thousand replications were conducted in each condition, each of which was as follows Twenty item easiness parameters were drawn from a normal distribution, $b \sim N(\mu_{easy,1})$, and the mean of that distribution was drawn from a continuous uniform distribution, $\mu_\theta \sim unif(-2,2)$. Similarly, 20 item discrimination parameters were drawn from a log-normal distribution, $a \sim Lognormal(\mu_a,0.5)$, and μ_a was drawn from a continuous uniform distribution, $\mu_a \sim unif(-0.5,1.5)$. Note that μ_a is the log of the median of the log-normal distribution so, for example, when $\mu_a = -0.5$, the expected median item discrimination is $exp(-0.5) \approx 0.61$. As in Simulation Study 1 and 2, for each replication, the 1PL, 2PL, and 3PL models were fit which determined the prediction-maximizing model according to ELPL-MR and ELPL-MP, and identified the model selected by BIC, AIC, and LRT.

Results for Simulation Study 3

Figure 3 shows the prediction-maximizing model for each replication according to ELPL-MR and ELPL-MP. As with Simulation Study 1 and 2, the 3PL model fit best more frequently according to ELPL-MP than ELPL-MR. The role of item easiness was as expected from Simulation Study 2: As μ_{easy} decreased, the 3PL model was more likely to win. Figure 4 shows the model selected by BIC, AIC, and LRT. Consistent with results from the previous two simulation studies, LRT selected models consistent with ELPL-MP, while BIC and AIC selected models more consistent with ELPL-MR.

As anticipated, the guessing parameter played a prominent role: The 3PL model usually won when $c = 0.25$, with the lowest sample size $I = 1000$ using ELPL-MR as an exception. Our original hypothesis was that $c = 0.03$ was a nearly ignorable level of guessing and consequently that the 3PL model would not perform well. That turned out not to be the case: The 3PL model won somewhat frequently even when $c = 0.03$. Turning to discrimination, as μ_a increased (so that overall item discrimination increased), the 2PL model performed worse. This result might seem counter-

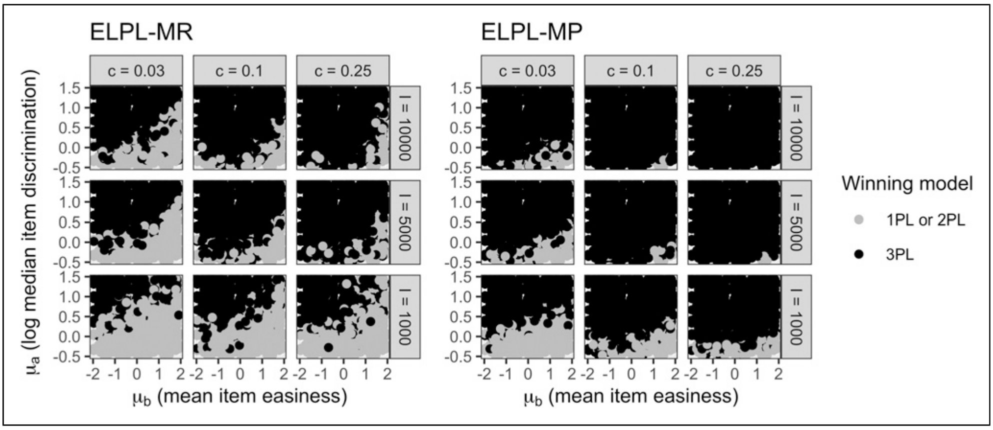


Figure 3. Simulation study 3 results for the predictive fit metrics, ELPL-MR and ELPL-MP. Each point corresponds to the prediction-maximizing model from one of 1000 replications. The 3PL model was most likely to offer the best fit with greater item discrimination, more difficult items, and more persons. ELPL-MP preferred the 3PL model more often than ELPL-MR did.

intuitive, but consider the following: For items with very high discriminations (i.e., nearly Guttman (1974) items), low-ability persons have very low probabilities of correct responses under the 2PL model, whereas in fact the true probability is never below the guessing parameter. This leads to the 2PL model performing poorly for items generated with high discrimination and some guessing.

Methods for Simulation Study

Each of the previous simulation studies looked at models with varying item complexity (e.g., 1PL, 2PL, and 3PL) but a fixed single latent ability factor (Figure 4). In Simulation Study 4, we invert our focus by always using a 2PL model, but varying the number of latent ability factors. For example, the 2-factor 2PL (hereafter 2F 2PL) model is specified as

$$Pr\left(Y_{ij}\right) = F\left(a_{j1}\theta_{i1} + a_{j2}\theta_{i2} + b_j\right)$$

where, for example, a_{j2} is the jth item's loading on the second factor, and θ_{i2} is the ith person's score for the second factor (Reckase, 2009). Our questions are similar as in the previous simulation studies: For example, if the DGM is a 2F 2PL model, when does a 1F 2PL model make better predictions for new data from the DGM as measured by ELPL-MR and ELPL-MP?

Accordingly, Simulation Study 4 used exclusively the 2F 2PL DGM. As with the previous simulation studies, only 20 items were considered. Two thousand runs were conducted, each of which was as follows (Figure 5) Item easiness parameters were drawn from the standard normal distribution, $b \sim N(0, 1)$. Item discrimination parameters were drawn independently from a log-normal distribution, $a \sim Lognormal(0, 0.5)$. The number of persons were drawn from a discrete uniform distribution, $I \sim unif\{500, 10000\}$. Abilities from a multidimensional normal distribution were drawn with mean vector $[\,\mu_{\theta_1} = 0, \mu_{\theta_2} = 0\,]$ and covariance matrix $\begin{bmatrix} 1 & v \\ v & 1 \end{bmatrix}$ where v is the correlation between factors and captures the degree to which persons with a high first factor score tend to have a high second factor score. For example, if the first factor is addition, and the second factor is subtraction, then v might be expected to be high. v can be thought as essentially making dimensionality continuous: At $v = 1$, ability is unidimensional, at $v = 0$, ability is fully two dimensional, and at $v = 0.5$, ability is somewhere between one and two dimensional. v is drawn from a continuous uniform distribution, $v \sim unif(0, 1)$. Two thousand such replications were conducted.

Results for Simulation Study 4

Figure 5 shows the winning model for each run according to ELPL-MP (left) and ELPL-MR (right). As before, ELPL-MR preferred more parsimonious models, with the 1F 2PL winning slightly more frequently according to ELPL-MR than ELPL-MP. The role of the correlation between factors is focused here, v. In general, as v increased, the 1F 2PL was more likely to win. As with Simulation Study 2 these results can be read in terms of minimum sample requirements for the 2F 2PL model. Under these conditions, the 2F 2PL was best according to both metrics whenever $v < 0.5$ (at least up to our minimum sample size of $I = 500$ persons). For greater values of v , the 1F 2PL was more often best, especially for lower sample sizes and according to ELPL-MR. Lastly, it is worth noting that the 2F 2PL typically won according to both metrics for v near 0.7 and I close to 10,000 persons, which suggests that at large sample sizes it's possible for multi-factor item response models to disentangle highly correlated factors.

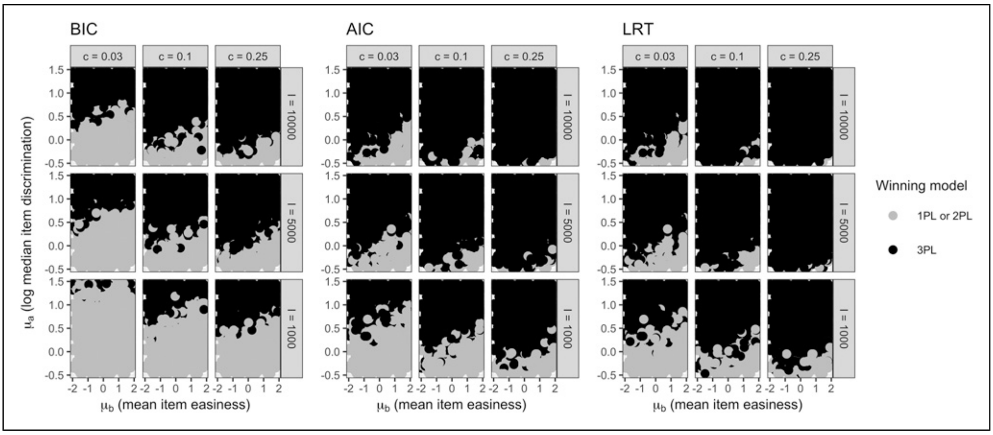


Figure 4. Simulation study 3 results for the model selection methods, BIC, AIC, and LRT. Each point corresponds to the selected model from one of 1000 replications. The 3PL model was selected more often with greater item discrimination, more difficult items, and more persons. BIC selected models consistent with ELPL-MR, while AIC and LRT selected models more consistent with ELPL-MP.

Predictive Fit in Practice via Cross-validation

In the simulation studies, ELPL-MR and ELPL-MP could be calculated because the DGM was known. In practice, when the DGM is not known, the predictive performance metrics can be estimated by hiding part of the data from the model so as to serve as out-of-sample data. This is known as cross-validation and it needs to be implemented based on which prediction task (and metric) is of interest (Bates et al., 2021). For example, Bolt and Lall (2003) implemented a cross-validation technique that corresponds to the missing person prediction task. Bergner et al. (2012) and Wu et al. (2020) cross-validated item response models in a way that corresponds to the missing responses task. How to use cross-validation to estimate each predictive fit metric in practice is now discussed. Then, a real data example of using cross-validation is provided.

Cross-validated Log Likelihood for Missing Responses (CVLL-MR)

The data are randomly partitioned into folds based on the item responses (i.e., the elements of Y). Randomization is stratified by person so that each person is split into “sub-persons” and each sub-person gets randomly assigned to a fold (DiTrapani, 2019). As a result, each fold contains approximately the same number of item responses for each person. Mathematically, following notation similar to Vehtari et al. (2017), the data is partitioned into eight folds $Y^{(k)}$ for $k = 1,\ldots,8$. Each model is fit separately to each training set $Y^{(-k)}$ using MMLE, which yields item parameter estimates $\boldsymbol{\psi}_{\mathbf{j}}^{(-\mathbf{k})}$. Similarly, person abilities, $\widehat{\boldsymbol{\theta}}_{\mathbf{i}}^{(-\mathbf{k})}$, are estimated using EAP. The predictive (i.e. out-of-sample or cross-validated) likelihood of $Y^{(k)}$ is

$$p\left(Y^{(k)}\middle|Y^{(-k)}\right)=\prod_{i=1}^{I}\prod_{j=1}^{J}\widehat{Pr}\left(y_{ij}^{(k)}\middle|\widehat{\boldsymbol{\psi}}_{\mathbf{j}}^{(-k)},\widehat{\boldsymbol{\theta}}_{\mathbf{i}}^{(-k)}\right).$$

Folds are then aggregated across to get the cross-validated log likelihood for missing responses (CVLL-MR)

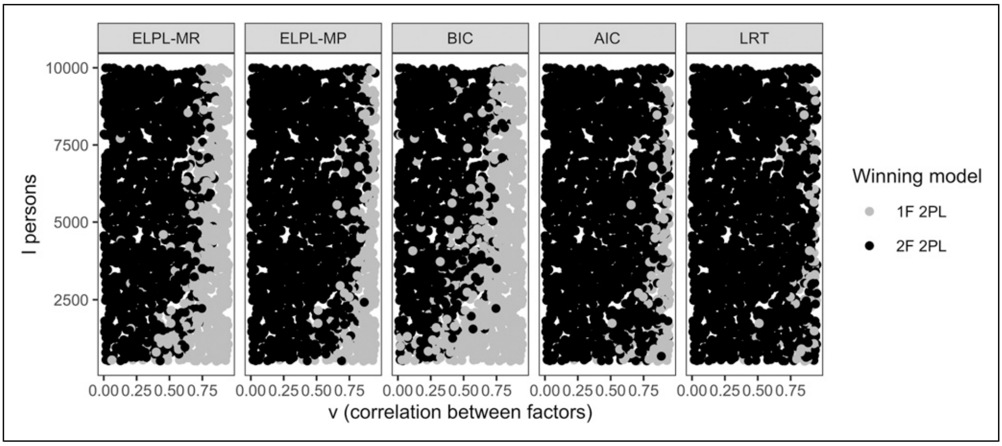


Figure 5. Simulation Study 4 results. Each point corresponds to the prediction-maximizing model according to the predictive fit metrics, ELPL-MR and ELPL-MP, or the model selected by BIC, AIC, or LRT for one of 2000 replications. The prediction-maximizing and selected model was more likely to be the 2F 2PL with lower correlation between factors and more persons. LRT nearly always selected the 2F 2PL model even in replications where both ELPL-MP and ELPL-MR identified the 1F 2PL model as prediction-maximizing. AIC selected models largely consistent with ELPL-MP. BIC selected models more closely aligned to ELPL-MR.

$$CVLL-MR(model(Y)) = \sum_{k=1}^{K} logp\left(Y^{(k)}\middle|Y^{(-k)}\right).$$

Cross-validated Log Likelihood for Missing Persons (CVLL-MP)

The data are randomly partitioned into folds based on the persons (i.e., the rows of Y). Mathematically, the data are partitioned into eight folds $Y^{(k)}$ for $k = 1,\ldots,8$. Each model is fit separately to each training set $Y^{(-k)}$ using MMLE, which yields item parameter estimates $\boldsymbol{\psi}_{\mathbf{j}}^{(-\mathbf{k})}$. For each fold, predictive (i.e. out-of-sample or cross-validated) likelihood of $Y^{(k)}$ is calculated by integrating over $\widehat{g}(\boldsymbol{\theta})$

$$p\left(Y^{(k)}\middle|Y^{(-k)}\right) = \prod_{i\in i^{(k)}}^{I} \int \prod_{j=1}^{J} \widehat{Pr}\left(y_{ij}^{(k)}\middle|\psi_j^{(-k)},\theta\right)\widehat{g}(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

Folds are then aggregated across to get the cross-validated log likelihood for missing persons (CVLL-MP)

$$CVLL-MP(model(Y)) = \sum_{k=1}^{K} logp\left(Y^{(k)}\middle|Y^{(-k)}\right).$$

Real Data Example

The Programme for International Student Assessment (PISA) is conducted every three years and aims to measure "the extent to which 15-year-old students, near the end of the compulsory education, have acquired key knowledge and skills" (PISA, 2015). In 2015, PISA switched from using a Rasch model to a 2PL model based on research showing that the 2PL model had lower AIC and BIC (Oliveri & Davier, 2011). To demonstrate model selection in practice, a variety of models fit was compared to a subset of the 2015 PISA data. In particular, the first 17 questions were considered focused (because they have high response rates) from science booklet 25. The 9800 students (regardless of country) who responded to all 17 questions were considered. As a result, Y contained no missingness and $9800 \cdot 17 = 166,600$ item responses. So that all items were dichotomously coded, partial credit was considered to be correct.

Seven models were compared that fit to this data: Rasch, 1F 2PL, 1F 3PL, 2F 2PL, 2F 3PL, 3F 2PL, and 3F 3PL. Each of these models was evaluated using CVLL-MR, CVLL-MP, BIC, AIC, and LRT. For the LRT, each pair of models ordered by number of parameters was compared. Each model was evaluated using cross-validated accuracy for missing responses (CVACC-MR), which is similar to CVLL-MR but accuracy (at a 0.5 cutoff) as opposed to log likelihood is aggregated across folds. As shown in Table 2, the selected model varied by metric. Consistent with results from our simulation studies, metrics based on the MR prediction task selected models with fewer parameters than metrics based on the MP prediction task. In particular, CVLL-MR and CVACC-MR both selected the 1F 3PL. The other four metrics, which are all based on marginalized likelihood, selected either the 2F 3PL or 3F 3PL models.

This case study demonstrates that the selected model varies significantly by metric. The cross-validation metrics involve fewer assumptions, which might be good reason to prefer them to the other metrics. Even so, which cross-validation metric should be used? If the goal is to predict probabilities of item responses—for example, in developing a computer adaptive version of the PISA—then the 1F 3PL model as selected by CVLL-MR and CVACC-MR[4] might be used. On the other hand, if the goal is draw conclusions with regard to the item parameters, then a model with more parameters such as the 3F 3PL model as selected by CVLL-MP may be preferred. This example illustrates that there is no single best fitting model—instead, results depend on a number of other factors including how prediction is defined.

Discussion

How should fit be thought about in the context of item response data? Previous research has frequently defined fit in terms of whether the model could have been the DGM (e.g., whether the expected contingency table from the model is similar to a contingency table of the data). An alternative view of fit is advocated, predictive fit, based on how well a model predicts new data from the DGM. Two predictive fit metrics are derived, ELPL-MR and ELPL-MP, which vary based on the meaning of out-of-sample for item responses. These metrics are derived in the theoretical case in which the DGM is a known item response model as is often the case in item response simulation studies and also provide an example of how they can be used in practice. As it is described below, it is believed that these metrics can help lay the groundwork for future advances in item response model evaluation in practice; are useful for evaluating item response models in simulation studies; and that our results offer guidance with regard to minimum sample size requirements for item response models.

The authors believe that predictive fit metrics can play a valuable role in laying the groundwork for future advances in item response model selection in practice. Model selection often involves comparing models with different numbers of parameters. A model with more parameters has a greater flexibility with which to fit the data-generating process but with this flexibility comes greater variance. A model with fewer parameters will have more stable estimates but with the stability comes potential bias. This is the well-known bias-variance trade off (Doroudi, 2020). Results from the simulation studies can be generalized in terms of this trade off. First, the prediction-maximizing model according to ELPL-MP was always more flexible than according to

ELPL-MR. In other words, the model that is best for the missing persons prediction task typically has more parameters than the model that is best for the missing responses prediction task. Second, AIC, BIC, and LRT varied across simulation studies in their ability to select the model that generated the best predictions. LRT selected the most flexible models, while BIC selected the least flexible models. LRT tended to select models that were even more flexible than ELPL-MP identified as prediction-maximizing, while BIC tended to select models that were even more conservative than ELPL-MR identified as prediction-maximizing.

In the real world example with PISA data, metrics based on the missing persons prediction task (CVLL-MP, AIC, BIC, and LRT) selected more flexible models than metrics based on the missing responses prediction task (CVLL-MR and CVACC-MR), which is generally consistent with results from the simulation studies. However, BIC selected a fairly flexible model (the 2F 3PL model) which is surprising given our simulation study results that showed BIC to be very conservative. One possible reason for this discrepancy is McDonald and Mok (1995)'s warning that AIC and BIC may fail with modest sample sizes or misspecified models. A solution is to use cross-validation which directly estimates the predictive fit metrics and requires fewer assumptions (Bates et al., 2021; Fang, 2011). Cross-validation was briefly described in our real world example but more research is needed to guide IRT practitioners in using cross-validation. For example: How many folds are necessary in k-fold cross-validation? How much better do estimates of the predictive fit metrics get as more folds are used?

This work offers more direct guidance on how models should be evaluated in simulation studies. For example, Kang and Cohen (2007) fit both a 2PL and 3PL model to data from a 3PL DGM. How should they have decided whether the 2PL model or the 3PL model fit better? They assumed that because the 3PL DGM was used that the 3PL model must fit better. Based on this assumption, they, for example, warned against using a model selection method because it frequently selected the 2PL model. An alternative is to consider predictive fit by determining which model makes the best predictions for additional data from the DGM. Results from our Simulation Study 1 demonstrate that in the conditions used in Kang and Cohen (2007), the 2PL model frequently actually makes better predictions than the 3PL model, and therefore has better predictive fit. Thus, it is a feature, not a bug, for a model selection method to select the 2PL model in these conditions. Our broader point is that predictive fit metrics should be considered in these types of simulation studies, and that using them has the potential to fundamentally change the study's conclusions.

Our simulation study results also offer guidance on a question of great practical importance: minimum sample size requirements for item response models. A variety of minimum sample size recommendations have been made for the 3PL model: Feuerstahler (2019) suggest at least 5000 persons, Hulin et al. (1982) suggest at least 1000 persons, and Thissen and Wainer (1982) suggest at least 100,000 persons. Despite these recommendations, Feuerstahler (2019) reports that "it is not uncommon to see the 3PL" model fit to item response data with fewer than 1000 persons (p. 12). In our view, a reasonable approach for benchmarking the minimum sample size for the 3PL model is to consider the sample size at which the 3PL model makes better predictions than the 2PL model. This is, of course, precisely what is investigated in the first three simulation studies. Our results indicate that the minimum sample size for the 3PL model depends on a variety of considerations, including how out-of-sample is defined, the ability of the persons, and the architecture of the items. For example, considering MR, greater average person ability, and greater item discrimination are all associated with the 3PL model producing relatively worse predictions, and thus greater minimum sample sizes for the 3PL model. Still, heuristics can be useful to practitioners: Simulation Study 2 results suggest a minimum sample size for the 3PL model of at least 1000 persons according to ELPL-MR and between 500 and 1000 persons according to ELPL-MP. With regard to multidimensional models, Simulation Study 4 results demonstrate that the minimum sample size requirement for the 2F 2PL model, defined by when the 2F 2PL model makes better predictions than the 1F 2PL model, depends greatly on the correlation between factors.

This study is closed with a fundamental question: How should item response models be evaluated and compared in practice? Should information criterion such as AIC and BIC, MP cross-validation where the empirical data is split at the response, or MR cross-validation where the data is split at the person be used? The authors believe that the answer likely depends on the purpose of the model. For example, the best model comparison method for selecting a model to identify poorly performing items might very well be different than that for selecting a model to rank-order persons. In the end, ELPL-MR and ELPL-MP are simply different ways of measuring the predictive performance of an item response model. High predictive performance is a desirable property for a model, but it certainly is not the only consideration (Vehtari et al., 2017). Future research might work to build a framework for comparing item response models in practice that helps researchers systematically balance the many considerations when navigating the process of starting with item response data, fitting a variety of models, and then selecting one of those models for a specific purpose.

Acknowledgments

The authors thank Klint Kanopka and Michael C. Frank for their invaluable feedback on drafts of this work.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was also supported by Institute of Education Sciences (grant number: R305B140009) and The Spencer Foundation Grant 20170008. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education.

ORCID iD

Benjamin A. Stenhaug https://orcid.org/0000-0001-5415-5540

Notes

[1] Both involve in-sample items. However, work by De Boeck (2008) proposes random item response models wherein out-of-sample items are tractable; future work could potentially focus on this case.

[2] Maximum likelihood ability estimates are not feasible because of completely perfect and imperfect response vectors. Future work might consider alternatives like weighted likelihood estimates (Warm, 1989).

[3] An alternative way to understand ELPL-MP preferring more flexible models is through the lens of regularization. Regularization typically counters over fitting by shrinking parameter estimates (Tibshirani, 1996). In this case, ELPL-MP treating ability as coming from $g(\boldsymbol{\theta})$ effectively regularizes the likelihood by which the model is judged. As a result, over fitting is punished less harshly.

[4] A benefit of missing responses is that the results are easier to interpret. For CVACC-MR, the accuracy of each model are all within 0.4% which may or may not be a small difference depending on the context. For CVLL-MR, the geometric mean of the likelihood can be reasoned about as the typical likelihood of an

individual item response: For the 1F 3PL model the geometric mean is $exp\left(\frac{-84828}{166600}\right) = 0.601$ as opposed to $exp\left(\frac{-86130}{166600}\right) = 0.596$ for the 1F Rasch model.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/tac.1974.1100705

Andersen, E. B. (1973). A goodness of fit test for the rasch model. *Psychometrika*, *38*(1), 123–140. https://doi.org/10.1007/bf02291180

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Bates, S., Hastie, T., & Tibshirani, R. (2021). Cross-validation: What does it estimate and how well does it do it? ArXiv Preprint arXiv:2104.00673.

Bergner, Y., Droschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., & Pritchard, D. E. (2012). *Model-based collaborative filtering analysis of student response data: Machine-learning item response theory.* International Educational Data Mining Society.

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Bock (1983). The discrete Bayesian. *Modern Advances in Psychometric Research*, 103–115. https://www.taylorfrancis.com/chapters/edit/10.4324/9780203056653-15/discrete-bayesian

Bock, & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/bf02293801

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using markov chain Monte Carlo. *Applied Psychological Measurement*, *27*(6), 395–414. https://doi.org/10.1177/0146621603258350

Casabianca, J. M., & Lewis, C. (2015). IRT item parameter recovery with marginal maximum likelihood estimation using loglinear smoothing models. *Journal of Educational and Behavioral Statistics*, *40*(6), 547–578. https://doi.org/10.3102/1076998615606112

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533. https://doi.org/10.1007/s11336-008-9092-x

DiTrapani, J. B. (2019). *Assessing the absolute and relative performance of IRTrees using cross-validation and the RORME index* (PhD thesis). The Ohio State University.

Doroudi, S. (2020). The bias-variance tradeoff: How data science can inform educational debates. *AERA Open*, *6*(4), 2332858420977208. https://doi.org/10.1177/2332858420977208

Embretson, S. E., & Reise, S. P. (2013). *Item response theory.* Psychology Press.

Fang, Y. (2011). Asymptotic equivalence between cross-validations and Akaike information criteria in mixed-effects models. *Journal of Data Science*, *9*(1), 15–21. https://doi.org/10.6339/JDS.201101_09(1).0002

Feuerstahler, L. (2019). Metric stability in item response models. Advance online publication. https://doi.org/10.1080/00273171.2020.1809980

Feynman, R. P., Leighton, R. B., & Sands, M. (1965). The Feynman lectures on physics; vol. I. *American Journal of Physics*, *33*(9), 750–752. https://doi.org/10.1119/1.1972241.

Furr, D. C. (2017). *Bayesian and frequentist cross-validation methods for explanatory item response models* (PhD thesis). Doctoral Dissertations.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2

Guttman, L. (1974). *The basis for scalogram analysis*. Bobbs-Merrill.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement*, *6*(3), 249–260. https://doi.org/10.1177/014662168200600301

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331–358. https://doi.org/10.1177/0146621606292213

Lord, F. M. (1983). Small n justifies Rasch model. In *New horizons in testing* (pp. 51–61). Elsevier.

Luecht, R., & Ackerman, T. A. (2018). A technical note on IRT simulation studies: Dealing with truth, estimates, observed data, and residuals. *Educational Measurement: Issues and Practice*, *37*(3), 65–76. https://doi.org/10.1111/emip.12185

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 71–101. https://doi.org/10.1080/15366367.2013.831680

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. https://doi.org/10.1198/016214504000002069

McDonald, R. P., & Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*(1), 23–40. https://doi.org/10.1207/s15327906mbr3001_2

Oliveri, M. E., & Davier, M. V. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*(3), 315. https://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/04_Oliveri.pdf

PISA, O. (2015). *PISA: Results in focus*. OECD: Organisation for Economic Co-Operation and Development.

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. ERIC.

Reckase, M. D. (2009). *Multidimensional item response theory models* (pp. 79–112). Springer. https://doi.org/10.1007/978-0-387-89976-3_4

Schwarz, G.others (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298–321. https://doi.org/10.1177/0146621605285517

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173–180. https://doi.org/10.1207/s15327906mbr2502_4

Svetina, D., & Levy, R. (2016). Dimensionality in compensatory MIRT when complex structure exists: Evaluation of DETECT and NOHARM. *The Journal of Experimental Education*, *84*(2), 398–420. https://doi.org/10.1080/00220973.2015.1048845

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*(4), 397–412. https://doi.org/10.1007/bf02293705

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/bf02294627

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? https://doi.org/10.31219/osf.io/u6vz5

Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Wu, M., Davis, R. L., Domingue, B. W., Piech, C., & Goodman, N. (2020). *Variational item response theory: Fast, accurate, and expressive*. ArXiv Preprint arXiv:2002.00276.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393