Predictive Fit Metrics for Item Response Models

Benjamin A. Stenhaug[1] & Benjamin W. Domingue[1]

[1] The Graduate School of Education at Stanford University

Author Note

Correspondence concerning this article should be addressed to Benjamin A. Stenhaug, 450 Serra Mall, Stanford, CA 94305. E-mail: stenhaug@stanford.edu

14   **Abstract**

15   The fit of an item response model is typically conceptualized as whether a given model could have

16   generated the data. We advocate for an alternative view of fit, "predictive fit,' ' based on the model's

17   ability to predict new data. We derive two predictive fit metrics for item response models that assess

18   how well an estimated item response model fits the data-generating model. These metrics are based

19   on long-run out-of-sample predictive performance (i.e., if the data-generating model produced

20   infinite amounts of data, what is the quality of a model's predictions on average?). The fundamental

21   difference between these metrics is the definition of out-of-sample on which they are built, which is

22   complicated by item responses being cross-classified within items and persons. We conduct

23   simulation studies to identify the prediction-maximizing model across a variety of conditions

24   including number of persons, average person ability, item discrimination, item guessability, and

25   correlation between multiple factors. In general, we show that considering persons to be

26   out-of-sample—as psychometricians often do—leads to a preference for more flexible models (i.e.,

27   those with more item parameters). In each simulation, we compare the prediction-maximizing

28   model to the model selected by AIC, BIC, and likelihood ratio tests. We find that performance of

29   these methods depends on the prediction task of interest. In general, likelihood ratio tests often

30   select overly flexible models, while BIC selects overly parsimonious models. We use PISA data to

31   demonstrate how to use cross-validation to directly estimate the predictive fit metrics in practice.

32   We discuss implications for item response model selection in operational settings.

33   *Keywords:* Item response theory; Fit; Prediction; Model comparison; Cross-validation

Predictive Fit Metrics for Item Response Models

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 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 is 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 (Albert 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.[1] 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, Johnson, & Stern, 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 *comparative model fit* script, wherein (1) a

---

[1] This comparison can also be translated into a goodness-of-fit metric such RMSEA$_2$ (Steiger, 1990).

DGM model is chosen (e.g., the 3PL), (2) item and person parameters are specified and item response data is 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 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 and others (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, a method for detecting the dimensionality of item response data, based on its tendency to find fewer than the data-generating number of dimensions at low sample sizes.

An Alternative Approach to Fit: Predictive Fit

We have summarized this previous work to illustrate how fit is often conceptualized in the item response theory research literature. An alternative approach, which has gained traction in computer science, is predictive fit which is based on how well a model predicts new (i.e., out-of-sample) data from the DGM (Gelman, Hwang, & Vehtari, 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 *applicable*" (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 both Lord and Rasch's observation is that the goal of model selection should be to identify a useful model—which can be measured by quality of predictions—not necessarily the "right" or "true" model.

We have summarized this previous work to illustrate how fit is often conceptualized in the item response theory research literature. An alternative approach, which has gained traction in computer science, is predictive fit (Gelman, Hwang, & Vehtari, 2014). The fundamental logic of predictive fit is that the model with the best predictions is likely to be the most useful. Box (1976) famously wrote that "all models are wrong" [p. 66]. Perhaps less famously, Rasch (1960), in the same text that introduced the Rasch model, wrote, "When you construct a model you leave out all the details... Models should not be true, but it is important that they are *applicable*" [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. Indeed, a compelling way to assess how applicable or useful an item response model is by the quality of its predictions. Following Gelman, Hwang, and Vehtari (2014), we define the predictive fit of an item response model by how well it predicts *new* data from the DGM. As is common, we refer to new data as out-of-sample data, as compared to in-sample data which was used to estimate the model's parameters.

One might argue that item response models are typically used to explain and understand as opposed to predict. We 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, Leighton, and Sands (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, we should not trust conclusions from item response models that make poor predictions. 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's 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, we argue that a compelling way 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 we name "missing responses," is to predict the probability of a missing item response. The second prediction task, which we name "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 we define as measures of how well an item response model predicts new data from the DGM.

Organization. We first give background on item response models and how they are typically compared in practice. Second, we describe the two possible prediction tasks which correspond to different definitions of out-of-sample for item response data. Third, we derive two predictive fit metrics based on these two definitions. We derive these metrics for the theoretical case when the DGM is known, such as in a simulation study. Fourth, we show the behavior and utility of these metrics in four simulation studies. To reexamine and extend Kang and Cohen (2007), we compare the prediction-maximizing model (according to the predictive fit metrics) to the model

selected selected by common methods such as AIC and BIC. Lastly, we provide a real world

example of model selection, 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. 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

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

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, $F(x) = \dfrac{e^x}{1 + e^x}$. 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.[2]

Item response models are most commonly estimated using marginal maximum likelihood

estimation (MMLE), which estimates item parameters while treating person abilities as a nuisance

parameter (Darreil 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 (Darrell Bock & Aitkin, 1981). As such, MMLE

---

[2] Typically, but not always, the specification of the IRF is the same for each item on an exam. For example, as is
common, we refer to the case where each of the items has a 3PL IRF as a 3PL model.

estimates item parameters by maximizing the marginal log likelihood (MLL) of the model,

$$\text{MLL}(\text{model}(Y)) = \sum_i \log \int \left[\prod_{j=1}^{J} \hat{\Pr}(y_{ij}|\hat{\psi}_j, \boldsymbol{\theta}) \right] \hat{g}(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2}$$

152   where $\hat{\psi}_j$ is the vector of estimated item parameters and $\hat{g}(\boldsymbol{\theta})$ is the distribution of ability. Typically

153   $\hat{g}(\boldsymbol{\theta})$ is assumed to be normally distributed with hyperparameters (e.g., mean and/or variance)

154   estimated during model fitting. [3] (Baker & Kim, 2004). If necessary, person ability estimates can

155   be obtained using an estimation technique such as expected a-posteriori (EAP) or maximum

156   a-posteriori (MAP) following MMLE of item parameters (Darreil Bock, 1983).

157   **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 (Alberto Maydeu-Olivares, 2013). Each of these

methods is based on the model's marginalized log likelihood, $\text{MLL}(\text{model}(Y))$, and number of

item parameters estimated in fitting the model, M. Consider two models fit to Y: $\text{model}_1(Y)$ and

$\text{model}_2(Y)$ 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

$$2 \cdot [\text{MLL}(\text{model}_2(Y)) - \text{MLL}(\text{model}_1(Y))] \tag{3}$$

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 $\text{model}_2(Y)$ fits better than $\text{model}_1(Y)$. On the other hand, AIC and BIC

---

[3] For example, the mirt R package assumes that $\hat{g}(\boldsymbol{\theta})$ follows a multivariate normal distribution by default. When fitting a 1PL model, the mean is fixed to 0 and the variance is estimated [@chalmers2012mirt]. When fitting a 2PL model, the mean is fixed to 0 and the variance is fixed to 1 (these fixed ability parameters are compensated for by free estimation of item difficulties and item discriminations, respectively).

both add a penalty to the likelihood based on M:

$$\text{AIC(model)} = -2 \cdot \text{MLL(model}(Y)) + 2M \tag{4}$$

$$\text{BIC(model)} = -2 \cdot \text{MLL(model}(Y)) + \log(I) \cdot M \tag{5}$$

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, we now take a step back to describe from first principles the missing responses and missing persons prediction tasks for item response models.

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 $\hat{g}(\boldsymbol{\theta})$. On the other hand, if it is single item responses that are out-of-sample, then we can use a person's responses to in-sample items to generate in-sample ability estimates.

We denote some arbitrary out-of-sample matrix $\tilde{Y}$. We consider two[4] versions of $\tilde{Y}$, 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. We denote this out-of-sample matrix as $\tilde{Y}^{\text{MR}}$, with "MR" abbreviating "missing responses." The unit of observation for $\tilde{Y}^{\text{MR}}$ 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

---

[4] 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.

177   ability. This logic can be applied to each entry $\tilde{Y}^{\text{MR}}$, and therefore $\tilde{Y}^{\text{MR}}$ has the same dimensions as

178   Y. Adaptive testing is an application in which the missing responses prediction task might make

179   sense: The goal of an adaptive testing engine is often to next assign an item that the person has a

180   fixed chance (e.g., 50%) of responding correctly to. Accordingly, the model that can best estimate

181   these probabilities may be most useful.

182   The second version of $\tilde{Y}$ comes from out-of-sample persons responding to in-sample items.

183   We denote this out-of-sample matrix as $\tilde{Y}^{\text{MP}}$, with `MP''` abbreviating missing persons.'' The

184   unit of observation for $\tilde{Y}^{\text{MP}}$ is a person's vector of item responses. The bottom row on the right of

185   Figure 1 represents a new person, Person D, responding to each of the items for the first time. The

186   prediction task is for the model to estimate the likelihood of all of Person D's item responses. We

187   can use the other persons to estimate item parameters, but we have no way to estimate Person D's

188   ability. As a result, we have to make a prediction about their entire vector of item responses—the

189   unit of analysis—by treating ability as a nuisance variable; to do this, we average (i.e., integrate)

190   over the distribution, denoted $g(\theta)$, from which we assume Person D's ability originates.[5] So that

191   $\tilde{Y}^{\text{MP}}$ has the same scale as Y, we might consider there to be as many missing persons as there are

192   persons in Y. Traditional linear testing is an application wherein the missing persons prediction task

193   might be preferred: the ability of the next test taker is unknown and a reasonable goal might be to

194   prefer a scoring model that will perform optimally in the case of this heightened uncertainty. Note

195   that LRT, AIC, and BIC all implicitly target the missing persons prediction task because they are

196   based on marginalized log likelihood of the model.

Predictive Fit Metrics

197   We now derive a predictive fit metric for each of $\tilde{Y}^{\text{MR}}$ and $\tilde{Y}^{\text{MP}}$. These metrics can only be

calculated exactly when the DGM is known such as in a simulation study (we later describe how to

use cross-validation to estimate these metrics in practice). When the DGM is known, we can

---

[5] This is how marginal maximum likelihood estimation (MMLE) treats ability when calculating likelihood (thus, "marginal" likelihood) [@baker2004item].

Figure 1

Understanding the two out-of-sample item response matrices, $\tilde{Y}^{MR}$ and $\tilde{Y}^{MP}$

directly measure a model's predictive performance on the *distribution of data* produced 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 we generically denote 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, Hwang, & Vehtari, 2014):

$$\text{lpl}(\tilde{Y}, \text{model}(Y)) = \log\hat{\Pr}(\tilde{Y}|\text{model}(Y)). \tag{6}$$

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

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

$$\text{lpl-MR}(\tilde{Y}^{MR}, \text{model}(Y)) = \log\hat{\Pr}(\tilde{Y}^{MR}|\text{model}(Y)) = \sum_{i=1}^{I}\sum_{j=1}^{J}\log\hat{\Pr}(\tilde{y}_{ij}^{MR}|\hat{\psi}_j, \hat{\theta}_i) \tag{7}$$

where $\tilde{y}_{ij}^{MR}$ is an item response from Y^{MR}, $\hat{\psi}_j$ is item j's vector of parameter estimates from model(Y), and $\hat{\theta}_i$ is person i's vector of ability estimates from model(Y). We use the short-hand

$\hat{p}_{i,j}^{\mathrm{MR}} = \hat{\mathrm{Pr}}(y_{i,j}^{\mathrm{MR}} = 1|\hat{\boldsymbol{\psi}}_j, \hat{\boldsymbol{\theta}}_i)$. To be concrete, in the case of the unidimensional 2PL model specification

$$\mathrm{lpl\text{-}MR}(\tilde{Y}^{\mathrm{MR}}, \mathrm{model}(Y)) = \tag{8}$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J} \tilde{y}_{ij}^{\mathrm{MR}} \log\left(\hat{p}_{i,j}^{\mathrm{MR}}\right) + (1 - \tilde{y}_{ij}^{\mathrm{MR}}) \log\left(1 - \hat{p}_{i,j}^{\mathrm{MR}}\right) = \tag{9}$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J} \tilde{y}_{ij}^{\mathrm{MR}} \log\left(F(\hat{a}_j\hat{\theta}_i + \hat{b}_j)\right) + (1 - \tilde{y}_{ij}^{\mathrm{MR}}) \log\left(1 - F(\hat{a}_j\hat{\theta}_i + \hat{b}_j)\right). \tag{10}$$

Of course, there are many possible out-of-sample item response matrices $\tilde{Y}^{\mathrm{MR}}$. The measure of model performance should be reflective of the DGM in general, not one particular $\tilde{Y}^{\mathrm{MR}}$. The DGM is captured simply by $p_{i,j}^{\mathrm{MR}}$, the data-generating probability of y_{ij}^{MR} being responded to correctly. If the DGM is an item response model, $p_{i,j}^{\mathrm{MR}} = \mathrm{Pr}(y_{i,j}^{\mathrm{MR}} = 1|\boldsymbol{\psi}_j, \boldsymbol{\theta}_i)$. 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:

$$\mathrm{ELPL\text{-}MR}(\mathrm{model}(Y)) = \mathbb{E}\left[\mathrm{lpl}(\tilde{Y}^{\mathrm{MR}}, \mathrm{model}(Y))\right] \tag{11}$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J} p_{i,j}\log(\hat{p}_{i,j}) + (1 - p_{i,j})\log(1 - \hat{p}_{i,j}). \tag{12}$$

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 $\tilde{Y}^{\mathrm{MR}}$.

One way to think about equation 12 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 $0.9\log 0.8 + 0.1\log 0.2 \approx -0.36$. Translating back to the probability scale, the long-run likelihood is $\exp(-0.36) \approx 0.70$.

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

We now derive the predictive fit metric for when the prediction task is the vector of responses for persons not known to the model as is a row vector, $\boldsymbol{y_u^{\text{MP}}}$, from $\tilde{Y}^{\text{MP}}$. Calculation of log predictive likelihood of missing persons (lpl-MP) for $\tilde{Y}^{\text{MP}}$ is complicated by the fact that the persons in $\tilde{Y}^{\text{MP}}$ 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 $\hat{g}(\boldsymbol{\theta})$, the distribution of ability as estimated by the model. We begin by calculating the lpl of $\boldsymbol{y_u^{\text{MP}}}$:

$$\text{lpl-MP}(\boldsymbol{y_u^{\text{MP}}}, \text{model}(Y)) = \int \hat{\text{Pr}}(\boldsymbol{y_u^{\text{MP}}}|\boldsymbol{\theta})\hat{g}(\boldsymbol{\theta})d\boldsymbol{\theta} = \int \left[\prod_{j=1}^{J} \hat{\text{Pr}}(y_{uj}^{\text{MP}})|\hat{\boldsymbol{\psi}}_{\boldsymbol{j}}, \boldsymbol{\theta}) \right] \hat{g}(\boldsymbol{\theta})d\boldsymbol{\theta} \qquad (13)$$

Next, we need to account for the data-generating distribution of $\tilde{Y}^{\text{MP}}$, 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:

$$\pi_u = \int \text{Pr}(\boldsymbol{y_u^{\text{MP}}}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta} = \int \left[\prod_{j=1}^{J} \text{Pr}(y_{uj}^{\text{MP}})|\boldsymbol{\psi}_{\boldsymbol{j}}, \boldsymbol{\theta}) \right] g(\boldsymbol{\theta})d\boldsymbol{\theta}. \qquad (14)$$

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 $\boldsymbol{y_u}$:

$$\text{ELPL-MP}(\text{model}(Y)) = \mathbb{E}\left[\text{lpl}(\boldsymbol{y_u^{\text{MP}}}, \text{model}(Y)) \right] \qquad (15)$$

$$= \sum_{u=1}^{U} \pi_u \cdot \text{lpl}(\boldsymbol{y_u^{\text{MP}}}, \text{model}(Y)) \qquad (16)$$

$$= \sum_{u=1}^{U} \pi_u \cdot \left(\int \left[\prod_{j=1}^{J} \hat{\text{Pr}}(y_{uj}|\hat{\boldsymbol{\psi}}_{\boldsymbol{j}}, \boldsymbol{\theta}) \right] \hat{g}(\boldsymbol{\theta})d\boldsymbol{\theta} \right) \qquad (17)$$

208 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 $\tilde{Y}^{\text{MP}}$. 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, we conducted four simulation studies. 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, we identified the prediction-maximizing model according to the two predictive

fit metrics. We also calculated which model AIC, BIC, and LRT 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 and ELPL-MP for each of

the models, and (4) determine which model AIC, BIC, and LRT selected. 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, we 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.

We used R for computing (R Core Team, 2019). We used the R package, mirt, to fit models

using MMLE with the EM algorithm and 61 quadrature points (Chalmers, 2012). We used custom

written functions to calculate ELPL-MR and ELPL-MP for each model. In particular, we calculated

ELPL-MR using equation 12. We estimated abilities using both MAP and EAP with the usual

234 standard normal prior. Because results using EAP and MAP were nearly identical, we report only

235 results using EAP ability estimates.[6] We calculated ELPL-MP using equation 17. Integrals were

236 approximated using Gauss-Hermite quadrature with 61 points (Embretson & Reise, 2013). We used

237 the suite of R packages known as the tidyverse for data wrangling and visualization (Wickham,

238 2017). Materials to reproduce this paper, including functions to estimate ELPL-MR and ELPL-MP,

239 are available at [blinded GitHub link].

Methods for Simulation Study 1

241   In Simulation Study 1, we revisited Kang and Cohen (2007) who evaluated model selection

242 methods (e.g., BIC) via their capacity to identify the model with the same parameterization as the

243 DGM (e.g., a model selection method should choose the 3PL model if the 3PL DGM was used).

244 We questioned whether the 3PL model actually had the best predictive fit in the conditions in which

245 they conducted their simulation study. We focused on the six conditions from Kang and Cohen

246 (2007) that came from crossing the DGM (1PL, 2PL, or 3PL) and sample size (500 or 1000

247 persons). In each condition, we used 20 items and drew abilities from a normal distribution,

248 $\theta \sim N(0,1)$. We used their exact item parameters as reported in Table 4 of Kang and Cohen

249 (2007).[7] We conducted 500 replications for each condition. Our hypothesis was that the winning

250 model—that is, the most predictive model—would not always have the same parameterization as

251 the DGM, especially at lower sample sizes.

Results for Simulation Study 1

253   Table 1 shows the number of replications in which each model won. In the conditions with

254 the 1PL or 2PL DGM, the prediction-maximizing model according to ELPL-MR and ELPL-MP

---

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

[7] These were generated in the original study by randomly drawing difficulties from a normal distribution, $b \sim N(0,1)$, discriminations from a log-normal distribution, $a \sim \text{Lognormal}(0,0.5)$, and guessing parameters from a beta distribution, $c \sim B(5,17)$. For the 1PL model, all discriminations were set to 1 and all guessing parameters were set to 0. For the 2PL, we all guessing parameters were set to 0.

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. We take this 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

we calculate ELPL-MR, we take the additional step of estimating each person's ability, which

causes the imperfection in the item parameter estimates to propagate to the person abilities. On the

other hand, when we calculate ELPL-MP, we just integrate over $g(\boldsymbol{\theta})$ which is much more tolerant

of those imperfect item parameter estimates.[8]

---

[8] 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, overfitting is punished less harshly.

Table 1

Simulation Study 1 results. Counts are of the winning model according to the theoretical predictive fit metrics, ELPL-MR and ELPL-MP, and the model selected by BIC, AIC, and LRT. With the 1PL and 2PL DGM, the predictive fit metrics find that the model with the same parameterization is the prediction-maximizing model and the model selection methods usually select this model. With the 3PL DGM, ELPL-MR tends to find that the 2PL model offers the best predictive fit, while ELPL-MP tends to find that the 3PL model offers the best predictive fit. LRT selects models consistent with ELPL-MP, while AIC and BIC selects models more consistent with ELPL-MR.

		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	18	0
2PL	500	0	499	1	0	499	1	0	500	0	0	500	0	0	500	0
2PL	1000	0	498	2	0	500	0	0	500	0	0	500	0	0	499	1
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

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. 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). We conducted 2000 replications, each of which was as follows. We drew the number of persons from a discrete uniform distribution, $I \sim \text{unif}\{100, 10000\}$. We drew abilities from a normal distribution, $\theta_i \sim N(\mu_\theta, 1)$, where the mean of that distribution was drawn from a continuous uniform distribution, $\mu_\theta \sim \text{unif}(-2, 2)$. As before, we simulated data 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 selected 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, we simulated item parameters with the goal of understanding how different item architectures—What is effect of discrimination? Of item easiness? Of guessing?—effect which model wins according to ELPL-MR and ELPL-MP. Simulation Study 3 again exclusively used the 3PL DGM. We first created nine conditions by crossing the vector of guessing parameters c (fixed to 0.03, 0.10, 0.25 for all items) and the sample size (set to 1000, 5000, or 10000 persons). We conducted 1000 replications in each condition, each of which was as follows. We drew 20 item easiness parameters from a normal distribution, $b \sim N(\mu_{\text{easy}}, 1)$, and we drew the mean of that
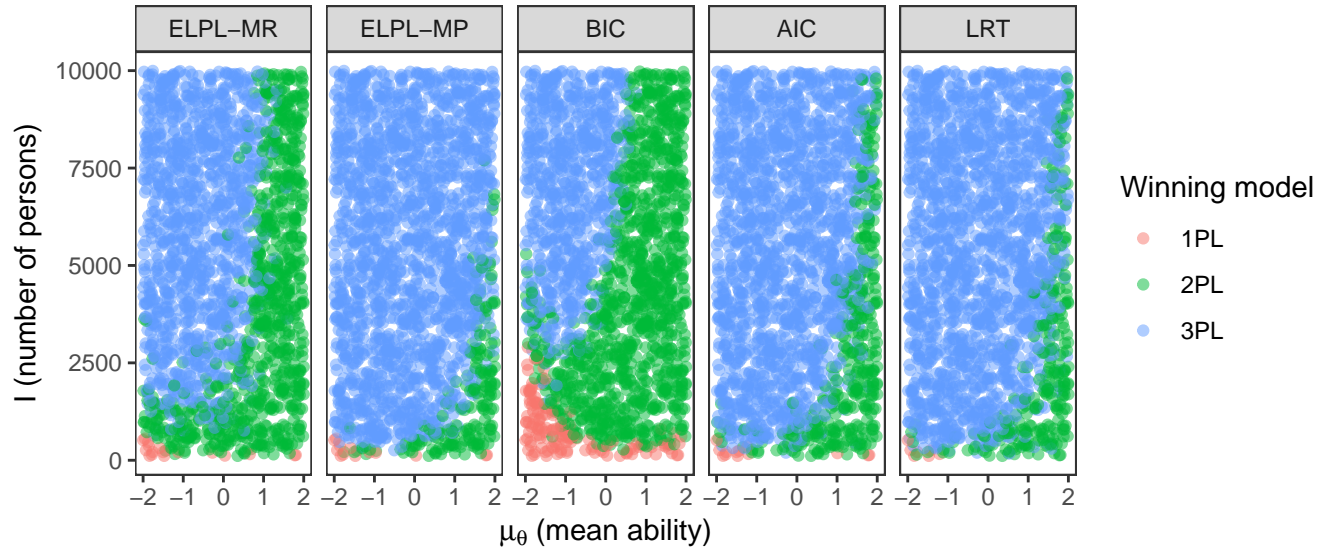
Figure 2

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.

316   distribution from a continuous uniform distribution, $\mu_\theta \sim \text{unif}(-2, 2)$. Similarly, we drew 20 item

317   discrimination parameters from a log-normal distribution, $a \sim \text{Lognormal}(\mu_a, 0.5)$, and we drew

318   μ_a from a continuous uniform distribution, $\mu_a \sim \text{unif}(-0.5, 1.5)$. Note that μ_a is the log of the

319   median of the log-normal distribution so, for example, when $\mu_a = -0.5$, the expected median item

320   discrimination is $\exp(-0.5) \approx 0.61$. As in Simulation Study 1 and 2, for each replication, we fit the

321   1PL, 2PL, and 3PL models determined the best fitting model according to ELPL-MR and

322   ELPL-MP, and identified the model selected by BIC, AIC, and LRT.

323   **Results for Simulation Study 3**

324   Figure 3 shows the prediction-maximizing model for each replication according to

325   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[9]

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-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 4

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. 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

$$\Pr(Y_{ij}) = F(a_{j1}\theta_{i1} + a_{j2}\theta_{i2} + b_j) \tag{18}$$

---

[9] In Simulation Study 2, the item easiness parameters were fixed and we varied the mean of ability. In Simulation Study 3, the ability distribution was fixed and we varied the mean of the item easiness parameters. The impact is the same: What matters is the difference between ability and item easiness.
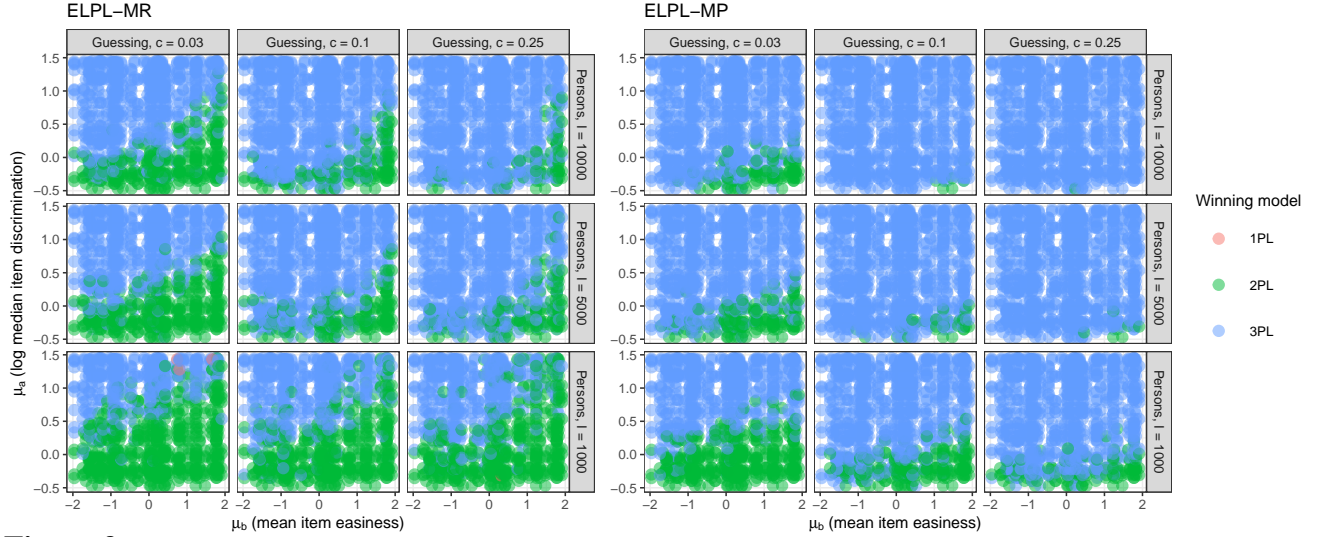
Figure 3

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.

343 where, for example, a_{j2} is the jth item's loading on the 2nd factor, and θ_{i2} is the ith person's score

344 for the 2nd factor (Reckase, 2009). Our questions are similar as in the previous simulation studies:

345 For example, if the DGM is a 2F 2PL model, does a 1F 2PL model or 2F 2PL model best fit the

346 DGM at a variety of sample sizes according to ELPL-MR and ELPL-MP?

347   Accordingly, Simulation Study 4 used exclusively the 2F 2PL DGM. As with the previous

348 simulation studies, we considered only 20 items. We conducted 2000 runs, each of which was as

349 follows. We drew item easiness parameters from the standard normal distribution, $b \sim N(0,1)$. We

350 drew item discrimination parameters independently[10] from a log-normal distribution,

351 $a \sim \text{Lognormal}(0,0.5)$. We drew the number of persons from a discrete uniform distribution,

352 $I \sim \text{unif}\{500,10000\}$. We drew abilities from a multidimensional normal distribution with mean

353 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

354 factors and captures the degree to which persons with a high first factor score tend to have a high

_____

[10] In particular, each item's loading on each factor was independent so that the first item's loading on the first factor was independent of its loading on the second factor.
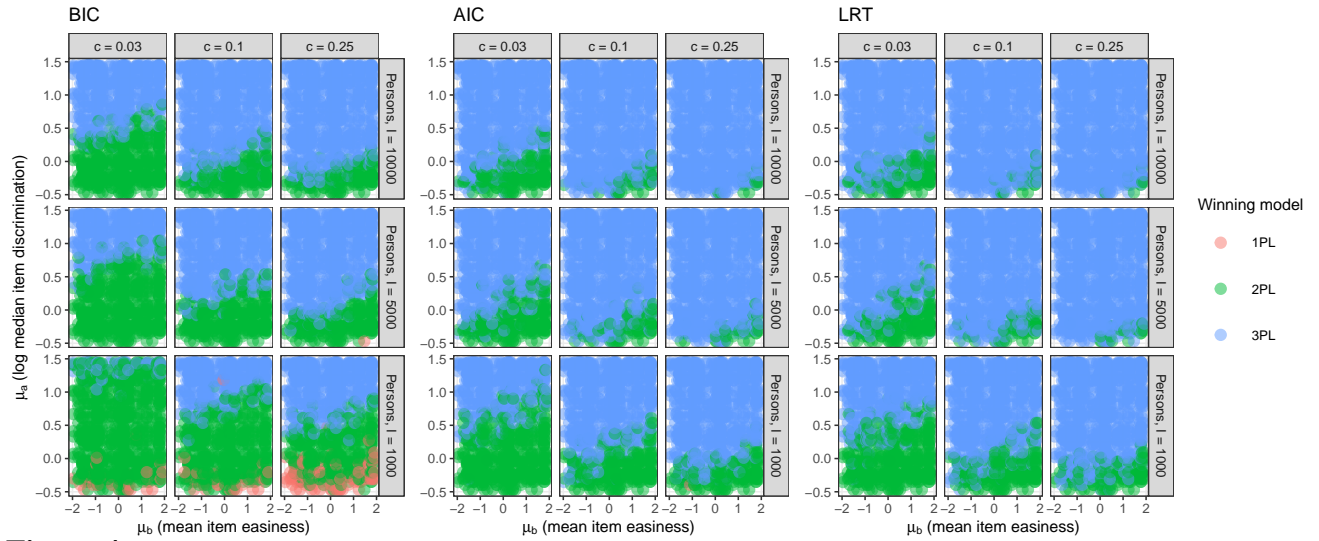
Figure 4

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.

355   second factor score. For example, if the first factor is addition, and the second factor is subtraction,

356   then we might expect v to be high. We can think of v as essentially making dimensionality

357   continuous: At $v = 1$, ability is unidimensional, at $v = 0$, ability is fully two-dimensional, and at

358   $v = 0.5$, ability is somewhere between one and two dimensional. We drew v from a continuous

359   uniform distribution, $v \sim \text{unif}(0, 1)$. We conducted 2000 such replications.

360   **Results for Simulation Study 4**

361   Figure 5 shows the winning model for each run according to ELPL-MP (left) and

362   ELPL-MR (right). As before, ELPL-MR preferred more parsimonious models, with the 1F 2PL

363   winning slightly more frequently according to ELPL-MR than ELPL-MP. We focus here on the role

364   of the correlation between factors, v. In general, as v increased, the 1F 2PL was more likely to win.

365   As with Simulation Study 2, we can read these results in terms of minimum sample requirements

366   for the 2F 2PL model. Under these conditions, the 2F 2PL was best according to both metrics

367   whenever $v < 0.5$ (at least up to our minimum sample size of $I = 500$ persons). For greater values

368   of v, the 1F 2PL was more often best, especially for lower sample sizes and according to ELPL-MR.

369 Lastly, it's worth noting that the 2F 2PL typically won according to both metrics for *v* near 0.7 and

370 *I* close to 10,000 persons, which suggests that at large sample sizes it's possible for multi-factor

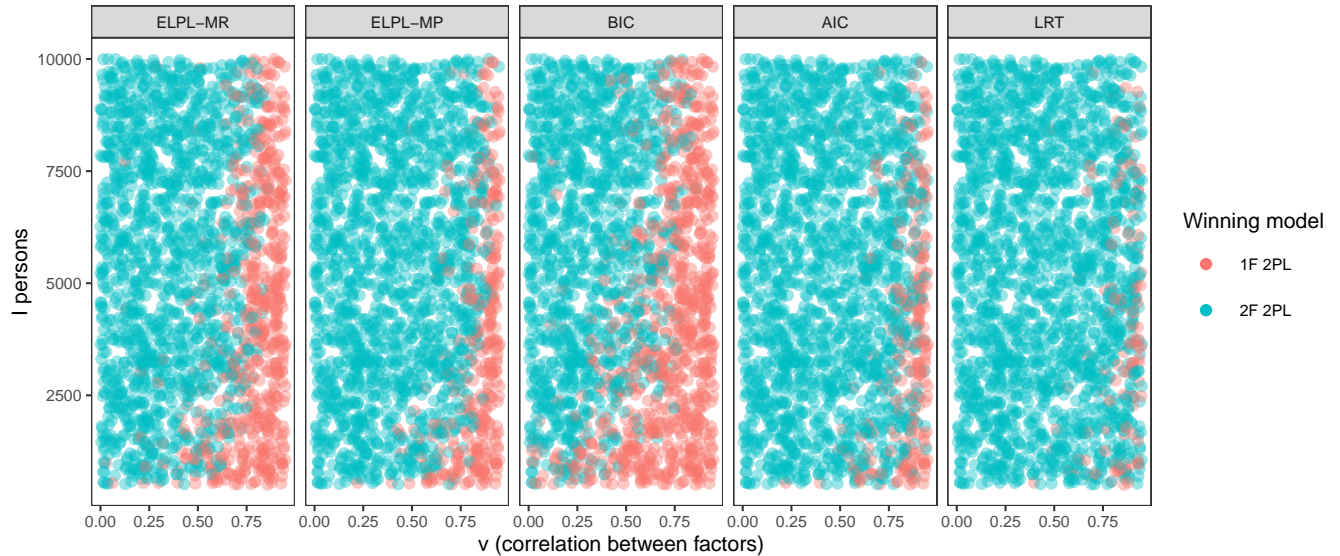371 item response models to disentangle highly correlated factors.

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.

Predictive Fit in Practice via Cross-validation

373 In the simulation studies, we could calculate ELPL-MR and ELPL-MP because the DGM

374 was known. In practice, when the DGM is not known, the predictive performance metrics can be

375 estimated by hiding part of the data from the model so as to serve as out-of-sample data. This is

376 known as cross-validation and it needs to be implemented based on which prediction task (and

377 metric) is of interest (Bates, Hastie, & Tibshirani, 2021). For example, Bolt and Lall (2003)

378 implemented a cross-validation technique that corresponds to the missing person prediction task.

379 Bergner et al. (2012) and Wu, Davis, Domingue, Piech, and Goodman (2020) cross-validated item

380 response models in a way that corresponds to the missing responses task. We now describe how to

381 use cross-validation to estimate each predictive fit metrics in practice. We then provide a real data

382 example of using cross-validation.

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

We randomly partition the data into folds based on the item responses (i.e., the elements of

Y). We stratify randomization 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, Gelman, and Gabry (2017), the data is partitioned into 8 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 $\hat{\psi}_j^{(-k)}$. Similarly, person abilities, $\hat{\theta}_i^{(-k)}$, are estimated using EAP. The

predictive (i.e. out-of-sample or cross-validated) likelihood of $Y^{(k)}$ is

$$p(Y^{(k)}|Y^{(-k)}) = \prod_{i=1}^{I} \prod_{j=1}^{J} \hat{\Pr}(y_{ij}^{(k)}|\hat{\psi}_j^{(-k)}, \hat{\theta}_i^{(-k)}). \tag{19}$$

We then aggregate across folds to get the cross-validated log likelihood for missing responses

(CVLL-MR):

$$\text{CVLL-MR}(\text{model}(Y)) = \sum_{k=1}^{K} \log p(Y^{(k)}|Y^{(-k)}). \tag{20}$$

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

We randomly partition the data into folds based on the persons (i.e., the rows of Y).

Mathematically, the data is partitioned into 8 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 $\psi_j^{(-k)}$. For each

fold, we calculate predictive (i.e. out-of-sample or cross-validated) likelihood of $Y^{(k)}$ by integrating

over $\hat{g}(\boldsymbol{\theta})$:

$$p(Y^{(k)}|Y^{(-k)}) = \prod_{i \in i^{(k)}} \int \prod_{j=1}^{J} \hat{\mathrm{Pr}}(y_{ij}^{(k)}|\psi_j^{(-k)}, \boldsymbol{\theta})\hat{g}(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{21}$$

We then aggregate across folds to get the cross-validated log likelihood for missing persons
(CVLL-MP):

$$\mathrm{CVLL\text{-}MP}(\mathrm{model}(Y)) = \sum_{k=1}^{K} \log p(Y^{(k)}|Y^{(-k)}). \tag{22}$$

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, we compared a
variety of models fit to a subset of the 2015 PISA data. In particular, we focused on the first 17
questions[11] from science booklet 25. We considered the 9800 students (regardless of country) who
responded to all 17 questions. As a result, Y contained no missingness and $9800 \cdot 17 = 166,600$
item responses. So that all items were dichotomously coded, we considered partial credit to be
correct.

We compared seven models fit to this data: Rasch, 1F 2PL, 1F 3PL, 2F 2PL, 2F 3PL, 3F
2PL, and 3F 3PL. We evaluated each of these models using CVLL-MR, CVLL-MP, BIC, AIC, and
LRT. For the LRT, we compared each pair of models ordered by number of parameters. We also
evaluated each model 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

―――――

[11] These questions have high response rates presumably because students don't run out of time

across folds. As shown in Table 1, 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 might be used. On the other hand, if the

goal is draw conclusions with regard to the population of persons, then a model with more

parameters such as the 3F 3PL model as selected by CVLL-MP may be preferred.

Model	Parameters	CVLL-MR	CVACC-MR	CVLL-MP	BIC	AIC	LRT
1F Rasch	18	-86130	0.74454	-89816	179759	179630	NA
1F 2PL	34	-84921	0.74708	-88844	177923	177679	<10^-10
2F 2PL	50	-85869	0.74757	-88758	177859	177499	<10^-10
1F 3PL	51	*-84828*	*0.74782*	-88756	177880	177513	<10^-10
3F 2PL	65	-86019	0.74764	-88741	177915	177448	<10^-10
2F 3PL	67	-85930	0.74734	-88682	*177837*	177355	<10^-10
3F 3PL	82	-86049	0.74736	*-88653*	177888	*177298*	*<10^-10*

\end{table}

Discussion

How should we think about fit 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). We advocate for an alternative view of fit, predictive fit, based on how well a model predicts new data from

the DGM. We derived two predictive fit metrics, ELPL-MR and ELPL-MP, which vary based on the meaning of out-of-sample for item responses. We derived these metrics 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 we describe below,

we believe that these predictive fit 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.

We 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

425   simulation studies can be generalized in terms of this tradeoff. First, the prediction-maximizing model according to ELPL-MP was always more flexible than according to ELPL-MR. In other

426   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

427   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.

428   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

429   identified as prediction-maximizing.

430   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

431   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)

432   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

433   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, Hastie, &

434   Tibshirani, 2021; Fang, 2011). We briefly described cross-validation in our real-world example but more research is needed to guide IRT practitioners in using cross-validation. For example: How

435   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?

436   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

437   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,

438   they, for example, warned against using a model selection method, BIC, in the conditions in which it frequently selected the 2PL model. An alternative is to consider predictive fit by determining

439   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

440   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 metric to select the 2PL model in these conditions. Our

441   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.

442   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

443   recommendations have been made for the 3PL model: Feuerstahler (2019) suggest at least 5000 persons, Hulin, Lissak, and Drasgow (1982) suggest at least 1000 persons, and Thissen and Wainer

444   (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

445   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

446   than the 2PL model. This is, of course, precisely what we investigate in the first three simulation studies. Our results indicate that the minimum sample size for the 3PL model depends on a variety

447   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

448   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

449   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

450   ELPL-MP. Turning for a moment to multidimensional models, Simulation Study 4 results demonstrate that the minimum sample size requirement for the 2F 2PL model, defined by when the 2F

451   2PL model makes better predictions than the 1F 2PL model, depends greatly on the correlation between factors.

452   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

453   recommendations have been made for the 3PL model: Feuerstahler (2019) suggest at least 5000 persons, Hulin, Lissak, and Drasgow (1982) suggest at least 1000 persons, and Thissen and Wainer

454   (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

455   persons [p. 12]. We believe that a reasonable way to think about the minimum sample size for the 3PL model is the sample size at which the 3PL model makes better predictions than the 2PL

456   model, which is precisely what the first three simulation studies investigated. Our results indicate that the minimum sample size for the 3PL model depends on a variety of considerations, including

457   how out-of-sample is defined, the ability of the persons, and the architecture of the items. For example, defining out-of-sample according to what we have called "missing responses," greater

458   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.

459   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

460   1000 persons according to ELPL-MP. 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

461   predictions than the 1F 2PL model, depends greatly on the correlation between factors.

462   We close 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

463   the empirical data is split at the *person level*, or MR cross-validation where the data is split at the *item response level* be used? We believe that the answer likely depends on the purpose of the

464   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.

465   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

466   model, but it certainly isn't the only consideration (Vehtari, Gelman, & Gabry, 2017). Future research might work to build a framework for comparing item response models in practice that helps

467   researchers systematically balance the many considerations when starting with item response data, fitting a variety of models, selecting one of those models, and then interpreting that model to

468   draw conclusions.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Andersen, E. B. (1973). A goodness of fit test for the rasch model. *Psychometrika*, *38*(1), 123–140.

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, Darreil. (1983). The discrete bayesian. *Modern Advances in Psychometric Research*, 103–115.

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

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.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

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.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, *48*(6), 1–29.

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

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.

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.

Feuerstahler, L. (2019). Metric stability in item response models.

Feynman, R. P., Leighton, R. B., & Sands, M. (1965). The feynman lectures on physics; vol. i. *American Journal of Physics*, *33*(9), 750–752.

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

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

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

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.

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

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

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

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

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

504   McDonald, R. P., & Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*(1), 23–40.

505   Oliveri, M. E., & Davier, M. von. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*(3),
506   315.

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

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

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

511   Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory* (pp. 79–112). Springer.

512   Schwarz, G., & others. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

513   Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298–321.

514   Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173–180.

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

517   Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*(4), 397–412.

518   Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

519   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.

520   Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

521   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?

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

523   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*.

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