# Optional ERIC Coversheet — Only for Use with U.S. Department of Education Grantee Submissions

This coversheet should be completed by grantees and added to the PDF of your submission if the information required in this form **is not included on the PDF to be submitted**.

**INSTRUCTIONS**

- Before beginning submission process, download this PDF coversheet if you will need to provide information not on the PDF.
- Fill in all fields—information in this form **must match** the information on the submitted PDF and add missing information.
- Attach completed coversheet to the PDF you will upload to ERIC [use Adobe Acrobat or other program to combine PDF files]—do not upload the coversheet as a separate document.
- Begin completing submission form at https://eric.ed.gov/submit/ and upload the full-text PDF with attached coversheet when indicated. Your full-text PDF will display in ERIC after the 12-month embargo period.

**GRANTEE SUBMISSION REQUIRED FIELDS**

**Title of article, paper, or other content**

Predictive Fit Metrics for Item Response Models

**All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.**


| Last Name, First Name | Academic/Organizational Affiliation | ORCID ID |
|---|---|---|
| Stenhaug, Ben | The Graduate School of Education at Stanford | |
| Domingue, Ben | The Graduate School of Education at Stanford | |
| | | |
| | | |
| | | |
| | | |


**Publication/Completion Date—**(if *In Press,* enter year accepted or completed) 2022

**Check type of content being submitted and complete one of the following in the box below:**

- ☑ If article: Name of journal, volume, and issue number if available
- ☐ If paper: Name of conference, date of conference, and place of conference
- ☐ If book chapter: Title of book, page range, publisher name and location
- ☐ If book: Publisher name and location
- ☐ If dissertation: Name of institution, type of degree, and department granting degree


Applied Psychological Measurement, 46(2), 136-155.

**DOI or URL to published work** (if available) https://doi.org/10.1177/01466216211066603


**Acknowledgement of Funding**— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.


"This work was supported by U.S. Department of Education **[Office name]** Institute of Education Scie through **[Grant number]** R305B1400 to **Institution]** Stanford Junio.The opinions expressed are those of the authors and do not represent views of the **[Office name]** Institute or the U.S. Department of Education.

# Predictive Fit Metrics for Item Response Models

Ben Stenhaug[1] & Ben Domingue[1]

[1] The Graduate School of Education at Stanford University

## Author Note

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B140009 to the Board of Trustees of the Leland Stanford Junior University. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education. The research reported here was also supported by The Spencer Foundation Grant 201700082. Thank you to Klint Kanopka and Michael C. Frank for their invaluable feedback on drafts of this work.

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

## Abstract

The fit of an item response model is typically conceptualized as whether a given model could have generated the data. We advocate for an alternative view of fit, “predictive fit”, based on the model’s ability to predict new data. We derive two predictive fit metrics for item response models that assess how well an estimated item response model (i.e., a data analysis 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 data analysis model’s predictions on average?). The fundamental difference between these metrics is the definition of out-of-sample on which they are built, which is complicated by item responses being cross-classified within items and persons. Via simulation studies, we show that (1) considering persons to be out-of-sample—as psychometricians often do—preferences more parsimonious models; (2) that when data is generated from a 3PL model, a 3PL data analysis model tends to make better predictions than a 2PL data analysis model with larger sample sizes, lower average ability, and lower average discrimination; and (3) that multidimensional models have better predictive fit when the correlation between ability factors is lower. We discuss implications for cross-validating item response models in practice.


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

Predictive Fit Metrics for Item Response Models

## Introduction

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 (i.e., data analysis models, abbreviated "DAM"). What does it mean for one of these DAMs to "fit" the data? In item response theory research literature, fit is most commonly defined by whether the DAM could have produced the data (DiTrapani, 2019). For example, the $M_2$ statistic compares the expected (according to the DAM) to the observed (by counting the data) moments of a contingency table (Maydeu-Olivares & Joe, 2005). Essentially, if these moments are similar enough, then we fail to reject lack of fit.[1] Similarly, posterior predictive checks use the DAM to simulate data, use discrepancy measures to compare that simulated data to the observed data, and then conclude whether the DAM 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 a great

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

many of these simulation studies as following the *comparative model fit* script, wherein (1) a 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 DAMs are estimated using the simulated data, and (4) those DAMs are compared. Luecht and Ackerman (2018) point out that inevitably it is concluded that the DAM 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 AIC and BIC. They simulated data via the 3PL DGM, and fit 1PL, 2PL, and 3PL DAMs 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 DAM as the best fitting model. Their implicit assumption was that, by definition, if a 3PL model generated the data, then a 3PL DAM must best fit the data. After all, no other model could have 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 used similar logic: 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 statistics and 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]. Indeed, a compelling way to assess how applicable or useful an item response model is by the quality of its predictions. Following Gelman et al. (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.

In operational settings, we cannot know that the data came from an item response model, but rather that 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 we can trust 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 DAM that produces high-quality predictions than it is to have a DAM with the same parameterization as the DGM. Accordingly, 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 DAM that makes the best predictions.

Predictive fit isn't how fit tends to be thought of in item response theory research literature, but it isn't new either: 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, psychometricians often compare item response models using information criterion such as AIC and BIC (Maydeu-Olivares, 2013). Information criterion essentially takes the predictive fit view: A penalty is added to the in-sample likelihood in order to be asymptotically equivalent to comparing models by predictive fit under a specific definition of the prediction task (Shao, 1997).

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" and is the view that information criterion for item response models usually takes, 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.

We focus on the theoretical case when the DGM is known, such as a simulation study. In this case, the predictive fit metrics can be calculated exactly. In particular, when the DGM is known, we can directly measure a DAM'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 DAM's fit to the DGM based on its predictive performance for this (infinite) out-of-sample data. Despite our focus on theortical conditions, we aim to lay the groundwork for future advances in item response model comparison methods in operational settings. 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. For example, Bolt and Lall (2003) implemented a cross-validation technique that corresponds to the missing person prediction task, and Bergner et al. (2012) cross-validated item response models in a way that corresponds to the missing responses task. We proceed by developing the two predictive performance metrics, but we return to the issue of model comparison in practice in the discussion.

**Organization.** We first describe the two possible prediction tasks which correspond to different definitions of out-of-sample for item response data. Second, we derive two predictive fit metrics based on these two definitions. Third, we show the behavior and utility of these metrics in four simulation studies. We close by discussing implications, including suggestions for model comparison in practice.

## Out-of-sample for Item Response Data

Let $Y$ represent an observed item response matrix. $y_{ij}$ is an observed dichotomous item response where $y_{ij} = 1$ indicates that the $i$th person responded correctly to the $j$th 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. 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 $\theta_i$ is the $i$th person's ability; $a_j$, $b_j$, and $c_j$ are the $j$th 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]

The goal of predictive fit metrics is to measure how well a DAM 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 responses that are 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. 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, but this fundamentally changes the measure by which we are evaluating a model's performance.

We denote some arbitrary out-of-sample matrix $\tilde{Y}$. We consider two[3] versions of $\tilde{Y}$, which

[2] Typically, but not always, the specification of the IRF is the Person Ae 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.

[3] 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.

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 $\tilde{Y}^{\mathrm{MR}}$, with “MR” abbreviating “Missing Responses”. The unit of observation for $\tilde{Y}^{\mathrm{MR}}$ 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 DAM’s prediction task is to estimate the probability of this missing response. To do so, the DAM can use the other persons to estimate the Item 1’s parameters and the other items to estimate Person A’s ability. This logic can be applied to each entry $\tilde{Y}^{\mathrm{MR}}$, and therefore $\tilde{Y}^{\mathrm{MR}}$ 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 is most useful.

The second version of $\tilde{Y}$ comes from out-of-sample persons responding to in-sample items. We denote this out-of-sample matrix as $\tilde{Y}^{\mathrm{MP}}$, with “MP” abbreviating “Missing Persons”. The unit of observation for $\tilde{Y}^{\mathrm{MP}}$ 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 DAM to estimate the likelihood of all of Person D’s item responses. We can use the other persons to estimate item parameters, but we have no way to estimate Person D’s ability. As a result, we have to make a prediction about their entire vector of item responses—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.[4] So that $\tilde{Y}^{\mathrm{MP}}$ has the same scale as $Y$, we might consider there to be as many missing persons as there are persons in $Y$. Traditional linear testing is an application in which the Missing Persons prediction task might make sense: It is unknown who will walk through the door to take the assessment next, and a reasonable goal might be to prefer a scoring model that can best estimate the probability of

[4] This is how marginal maximum likelihood estimation (MMLE) treats ability when calculating likelihood (thus, “marginal” likelihood) (Baker & Kim, 2004).

their string of item responses.

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

| | Item 1 | Item 2 | Item 3 |
|---|---|---|---|
| **Person A** | **?** | 0 | 1 |
| **Person B** | 0 | 1 | 0 |
| **Person C** | 1 | 1 | 0 |

$\tilde{Y}^{\mathrm{MP}}$ **(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, $\tilde{Y}^{\mathrm{MR}}$ and $\tilde{Y}^{\mathrm{MP}}$

## Predictive Fit Metrics

We now derive a predictive fit metric for each of $\tilde{Y}^{\mathrm{MR}}$ and $\tilde{Y}^{\mathrm{MP}}$. In general, both metrics measure how well a DAM 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 $\mathrm{model}(Y)$. 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 of the likelihood of $\tilde{Y}$ according to that model (Gelman et al., 2014):

$$\mathrm{lpl}(\tilde{Y}, \mathrm{model}(Y)) = \log \hat{\Pr}(\tilde{Y} | \mathrm{model}(Y)). \tag{2}$$

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

Calculation of lpl for $\tilde{Y}^{\mathrm{MR}}$ is relatively straightforward because we can use estimates of person abilities so that

$$\mathrm{lpl}(\tilde{Y}^{\mathrm{MR}}, \mathrm{model}(Y)) = \log \hat{\Pr}(\tilde{Y}^{\mathrm{MR}} | \mathrm{model}(Y)) = \sum_{i=1}^{I} \sum_{j=1}^{J} \log \hat{\Pr}(\tilde{y}_{ij} | \hat{\boldsymbol{\psi}}_j, \hat{\boldsymbol{\theta}}_i) \tag{3}$$

where $y_{ij}$ is an item response from $Y^{\mathrm{MR}}$, $\hat{\boldsymbol{\psi}}_j$ is item $j$'s vector of parameter estimates from model$(Y)$, and $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}$ is person $i$'s vector of ability estimates from model$(Y)$. The most common item response model estimation method, marginal maximum likelihood estimation (MMLE), does not directly provide ability estimates, but these are easily obtained using an estimation technique such as expected a-posteriori (EAP) or maximum a-posteriori (MAP) estimation after item parameters are estimated (Bock, 1983; Casabianca & Lewis, 2015)]. To be concrete, in the case of the dichotomous unidimensional 2PL model specification

$$\text{lpl}(\tilde{Y}^{\text{MR}}, \text{model}(Y)) = \sum_{i=1}^{I}\sum_{j=1}^{J} \tilde{y}_{ij} \log\left(F(\hat{a}_j\hat{\theta}_i + \hat{b}_j)\right) + (1-\tilde{y}_{ij})\log\left(1 - F(\hat{a}_j\hat{\theta}_i + \hat{b}_j)\right). \quad (4)$$

Of course, there are many possible out-of-sample item response matrices $\tilde{Y}^{\text{MR}}$. The measure of model performance should be reflective of the true DGM in general, not one particular $\tilde{Y}^{\text{MR}}$. Let $f(\tilde{Y}^{\text{MR}})$ represent the data-generating distribution of $\tilde{Y}^{\text{MR}}$. When the DGM is an item response model, $f(\tilde{Y}^{\text{MR}})$ includes the data-generating parameters for each item, $\boldsymbol{\psi}_{\boldsymbol{j}}$, and the data-generating abilities for each person, $\boldsymbol{\theta}_{\boldsymbol{i}}$. 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 taken over $f(\tilde{Y}^{\text{MR}})$:

$$\text{ELPL-MR}(\text{model(Y)}) = \mathbb{E}\left[\text{lpl}(\tilde{Y}^{\text{MR}}, \text{model}(Y)\right] = \int \sum_{i=1}^{I}\sum_{j=1}^{J} \log \hat{\Pr}(\tilde{y}_{ij}|\hat{\boldsymbol{\psi}}_{\boldsymbol{j}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{i}}) f(\tilde{Y}^{\text{MR}}) d\tilde{Y}^{\text{MR}}. \quad (5)$$

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

Ultimately, $f(\tilde{Y}^{\text{MR}})$ determines the data-generating probability of each item response. In the dichotomous case, let $\pi_{i,j}$ represent the true data-generating probability of the $i$th person responding correctly to the $j$th item. Similarly, let $\hat{\pi}_{i,j}$ represent the probability of the $i$th person responding correctly to the $j$th item as estimated by the DAM.[5] ELPL-MR can then be reduced to

[5] That is, $\hat{\pi}_{i,j} = \Pr(y_{i,j} = 1|\hat{\boldsymbol{\psi}}_{\boldsymbol{j}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{i}})$ and when the DGM is an item response model, $\pi_{i,j} = \Pr(y_{i,j} = 1|\boldsymbol{\psi}_{\boldsymbol{j}}, \boldsymbol{\theta}_{\boldsymbol{i}})$.

$$\text{ELPL-MR}(\text{model(Y)}) = \sum_{i=1}^{I}\sum_{j=1}^{J} \pi_{i,j}\log(\hat{\pi}_{i,j}) + (1-\pi_{i,j})\log(1-\hat{\pi}_{i,j}). \tag{6}$$

One way to think about equation 6 is that ELPL-MR is the weighted average of the log likelihood, where the weights are determined by the true probabilities. As an example, consider a DAM 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 DAM 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 $\exp(-0.36) \approx 0.70$.

## 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, $\boldsymbol{y_u}$, from $\tilde{Y}^{\text{MP}}$. Calculation of lpl 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 $\hat{g}(\theta)$, the distribution of ability as estimated by the DAM[6] (Baker & Kim, 2004). We begin by calculating the lpl of $\boldsymbol{y_u}$:

$$\text{lpl}(\boldsymbol{y_u}, \text{model}(Y)) = \int \hat{\Pr}(\boldsymbol{y_u}|\theta)\hat{g}(\theta)d\theta = \int \left[\prod_{j=1}^{J} \hat{\Pr}(y_{uj})|\hat{\boldsymbol{\psi}}_{\boldsymbol{j}}, \theta)\right]\hat{g}(\theta)d\theta \tag{7}$$

Next, we need to account for the data-generating distribution of $\tilde{Y}^{\text{MP}}$, which is captured by $\pi_u$, 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 $\pi_u$ as follows:

[6] For example, the mirt R package assumes that $g(\theta)$ follows a normal distribution by default. When fitting a 1PL model, the mean is fixed to 0 and the variance is estimated (Chalmers, 2012). 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).

$$\pi_u = \int \Pr(\boldsymbol{y_u}|\theta)g(\theta)d\theta = \int \left[\prod_{j=1}^{J} \Pr(y_{uj})|\boldsymbol{\psi_j},\theta)\right] g(\theta)d\theta. \tag{8}$$

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 $\boldsymbol{y_u}$:

$$\text{ELPL-MP}(\text{model}(Y)) = \mathbb{E}\left[\text{lpl}(\boldsymbol{y_u}, \text{model}(Y))\right] = \sum_{u=1}^{U} \pi_u \cdot \text{lpl}(\boldsymbol{y_u}, \text{model}(Y)) \tag{9}$$

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 $\tilde{Y}^{\text{MP}}$. Putting it all together, we arrive at

$$\textbf{??}\text{ELPL-MP}(\text{model}(Y)) = \sum_{u=1}^{U} \left(\int \left[\prod_{j=1}^{J} \Pr(y_{uj})|\boldsymbol{\psi_j},\theta)\right] g(\theta)d\theta\right) \left(\int \left[\prod_{j=1}^{J} \hat{\Pr}(y_{uj})|\hat{\boldsymbol{\psi}}_{\boldsymbol{j}},\theta)\right] \hat{g}(\theta)d\theta\right). \tag{10}$$

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 DAM 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.

In each of the simulation studies, we used R for computing (R Core Team, 2019). We used the R package, mirt, to fit DAMs 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 DAM. In particular, we calculated ELPL-MR using equation 6. We estimated abilities using both MAP and EAP with the usual standard normal prior. Because results using EAP and MAP were nearly identical, we report only results using EAP ability estimates.[7] We calculated ELPL-MP 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 paper, including functions to estimate ELPL-MR and ELPL-MP, are available at [blinded GitHub link].

## Methods for Simulation Study 1

In Simulation Study 1, we revisited Kang and Cohen (2007) who evaluated model selection methods (e.g., BIC) via their capacity to identify the DAM with the same parameterization as the DGM (e.g., a model selection method should choose the 3PL DAM if the 3PL DGM was used). We wondered whether the 3PL DAM actually had the best predictive fit in the conditions in which they conducted their simulation study. We focused on 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). In each condition, we used 20 items and drew abilities from a normal distribution, $\theta \sim N(0,1)$. We used their exact item parameters as reported in Table 4 of Kang and Cohen (2007).[8] For the 1PL DAM, we set all discriminations to 1 and all guessing parameters to 0. For the 2PL, we set all guessing parameters to 0. We conducted 500 “replications” for each condition. A replication consisted of the following steps: Simulate data using the DGM; fit a 1PL, 2PL, and 3PL DAM to the simulated data; and calculate the predictive performance metrics, ELPL-MR and ELPL-MP, for each of the DAMs. We considered the best-fitting model (i.e., the winning model) to be that which

[7] 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).

[8] 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, $c \sim B(5,17)$.

has the maximum ELPL-MR or ELPL-MP value. Our hypothesis was that the winning DAM would not always have the same parameterization as the DGM, especially at lower sample sizes.

## Results for Simulation Study 1

Table 1 shows the number of replications in which each DAM won according to both ELPL-MR and ELPL-MP. In the conditions with the 1PL or 2PL DGM, the winning DAM always shared the DGM's parameterization. However, for the 3PL DGM, the 2PL DAM often won (i.e., optimally predicted the out-of-sample data). For example, with a 3PL DGM and 500 persons, the 2PL DAM outperformed the 3PL DAM in 486 out of 500 replications according to ELPL-MR but only in 276 out of 500 replications according to ELPL-MP. Under these conditions, Kang and Cohen (2007) found that AIC selected the 2PL DAM in 96% if runs, which they interpreted as a failure of AIC. Our results show that if the goal is to find the model with the greatest ELPL-MR, AIC may actually have been performing quite well.

The model with the greatest ELPL-MR was often simpler than the model with the greatest ELPL-MP, which we take 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, $g(\theta)$, 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 $g(\theta)$ which is much more tolerant of those imperfect item parameter estimates.[9]

Table 1

*Simulation Study 1 results. We conducted 500 runs and calculated the winning DAM according to each of ELPL-MR and ELPL-MP.*

| | | **ELPL-MR** | | | **ELPL-MP** | | |
|---|---|---|---|---|---|---|---|
| DGM | Persons | 1PL | 2PL | 3PL | 1PL | 2PL | 2PL |
| 1PL | 500 | 500 | 0 | 0 | 500 | 0 | 0 |
| 1PL | 1000 | 500 | 0 | 0 | 500 | 0 | 0 |
| 2PL | 500 | 0 | 500 | 0 | 0 | 500 | 0 |
| 2PL | 1000 | 0 | 498 | 2 | 0 | 500 | 0 |
| 3PL | 500 | 0 | 486 | 14 | 0 | 276 | 224 |
| 3PL | 1000 | 0 | 405 | 95 | 0 | 9 | 491 |

## Methods for Simulation Study 2

Simulation Study 1 showed that with the 3PL DGM, the 2PL DAM is frequently best according to predictive performance metrics, especially if the number of persons is 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 DAM 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

[9] 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 $\hat{g}(\theta)$ effectively regularizes the likelihood by which the model is judged. As a result, overfitting is punished less harshly.

continuous uniform distribution, $\mu_\theta \sim \text{unif}(-2,2)$. As before, we simulated data using these parameters, fit the 1PL, 2PL, and 3PL DAMs, and determined the best fitting model according to ELPL-MR and ELPL-MP.

## Results for Simulation Study 2

Figure 2 shows the winning DAM for each replication according to ELPL-MP (left) and ELPL-MR (right). As in Simulation Study 1, ELPL-MR preferred more parsimonious models as evidenced by the 2PL DAM 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 DAM was to win. However, the ability distribution is also salient. As $\mu_\theta$ 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 DAM including a guessing parameter.

Albeit for a specific set of item parameters, Figure 2 can be read in terms of minimum sample requirements for the 3PL DAM. When $\mu_\theta$ is less than 0, the sample size at which the 3PL DAM tended to outperform the 2PL DAM was somewhat low ($\approx$ 2000) according to ELPL-MP, and it was a bit higher according to ELPL-MR. As $\mu_\theta$ increased, the relative predictive performance of the 3PL DAM decreased quickly, so much so that, for ELPL-MR, the 3PL DAM nearly never won when $\mu_\theta$ 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 effect which DAM wins according to ELPL-MR and ELPL-MP. What effect does greater item discrimination have? What about item easiness? And what role do different magnitudes of guessing behavior play?

Simulation Study 3 again exclusively used the 3PL DGM. We first created nine conditions corresponding to crossing the vector of guessing parameters $c$ (fixed to 0.03, 0.10, 0.25 for all

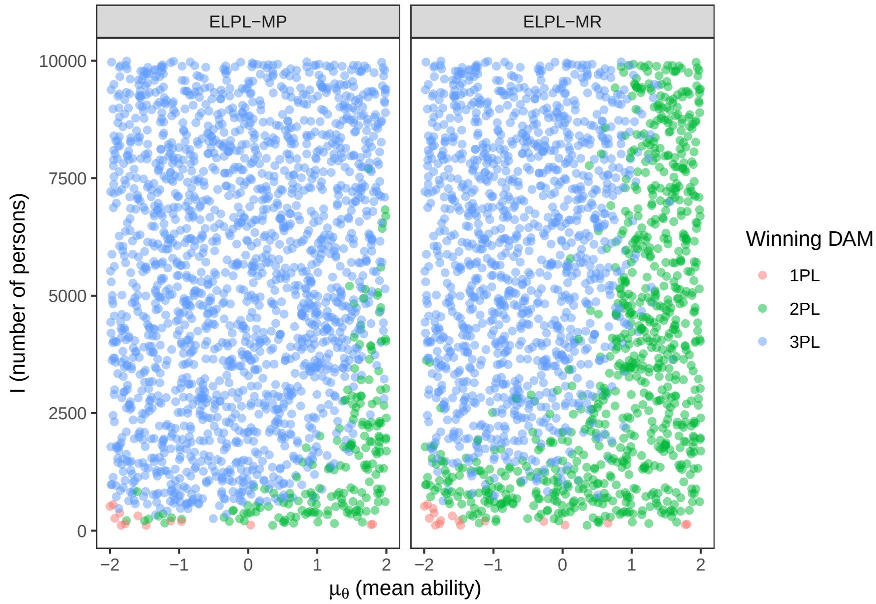


*Figure 2*. Simulation Study 2 results. Each point corresponds to the winning DAM from one of 2000 replications according to each of ELPL-MP (left) and ELPL-MR (right). A replication consisted of (1) simulating item response data with a random number of persons and a random mean ability; (2) fitting a 1PL, 2PL, 3PL DAM to that data; and (3) determining the best DAM according to ELPL-MP and ELPL-MR.

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 distribution from a continuous uniform distribution, $\mu_\theta \sim \text{unif}(-2,2)$. Similarly, we drew 20 item discrimination parameters from a log-normal distribution, $a \sim \text{Lognormal}(\mu_a, 0.5)$, and we drew $\mu_a$ from a continuous uniform distribution, $\mu_a \sim \text{unif}(-0.5, 1.5)$. Note that $\mu_a$ 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, we fit the 1PL, 2PL, and 3PL DAMs and then determined the best fitting model according to ELPL-MR and ELPL-MP.

## Results for Simulation Study 3

Figure 3 shows the winning DAM for each replication according to ELPL-MR. Figure 4 shows the same for ELPL-MP. As with Simulation Study 1 and 2, the 3PL DAM won more frequently according to ELPL-MP than ELPL-MR. The role of item easiness was as expected[10] from Simulation Study 2: As $\mu_{\text{easy}}$ decreased, the more likely the 3PL DAM was to win.

As anticipated, the guessing parameter played a prominent role: The 3PL DAM 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 nearly no guessing and consequently the 3PL DAM would not perform well. That turned out not to be the case: The 3PL DAM won somewhat frequently even when $c = 0.03$. Turning to discrimination, as $\mu_a$ increased (so that overall item discrimination increased), the 2PL DAM performed worse. Although counter-intuitive, 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 without a guessing parameter.

## 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

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

where, for example, $a_{j2}$ is the $j$th item's loading on the 2rd factor, and $\theta_{i2}$ is the $i$th person's score for the 2nd factor. Our questions are similar as in the previous simulation studies: For example, if

[10] 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.

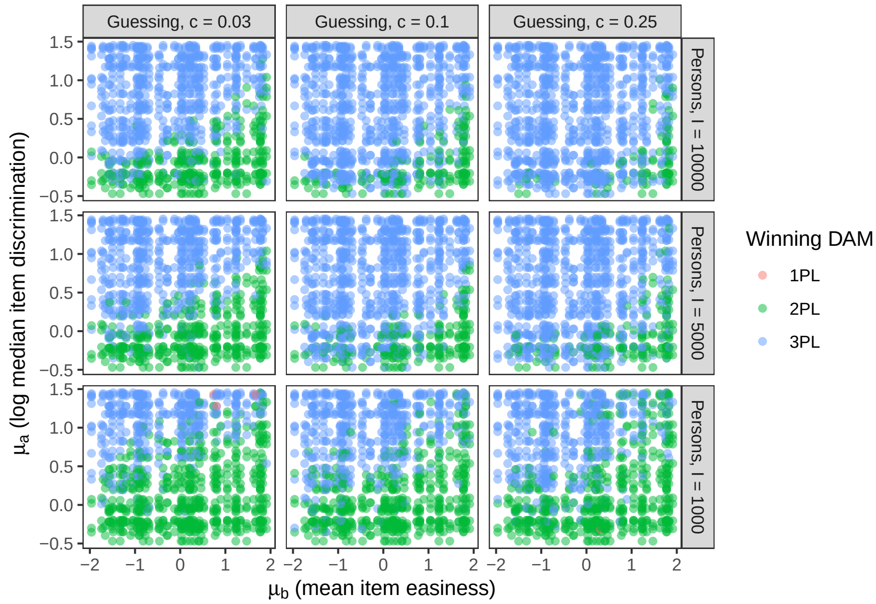


*Figure 3*. Simulation Study 3 results for ELPL-MR. Each point in each cell corresponds to the winning DAM from one of 1000 replications according to ELPL-MR. A replication consisted of (1) simulating item response data with a fixed guessing, a random mean item easiness, and a random log median item discrimination; (2) fitting a 1PL, 2PL, 3PL DAM to that data, and (3) determining the best DAM according to ELPL-MR.

the DGM is a 2F 2PL model, does a 1F 2PL model or 2F 2PL model best fit the DGM at a variety of sample sizes according to ELPL-MR and ELPL-MP?

Accordingly, Simulation Study 4 used exclusively the 2F 2PL DGM. As with the previous simulation studies, we consider only 20 items. We conducted 2000 runs, each of which was as follows. We drew item easiness parameters from the standard normal distribution, $b \sim N(0,1)$. We drew item discrimination parameters independently[11] from a log-normal distribution, $a \sim \text{Lognormal}(0, 0.5)$. We drew the number of persons from a discrete uniform distribution, $I \sim \text{unif}\{500, 10000\}$. We drew abilities from a multidimensional normal distribution with mean

[11] 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.

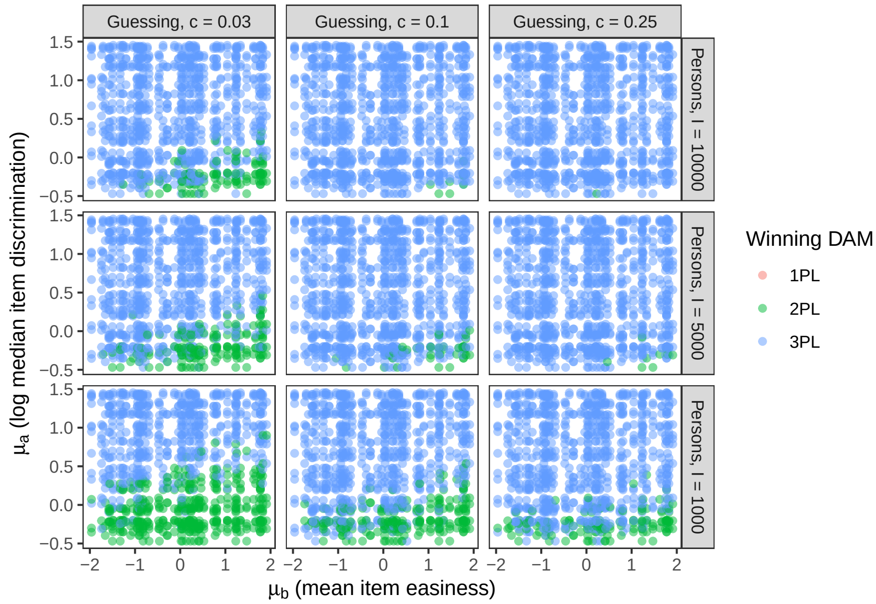


*Figure 4*. Simulation Study 3 results for ELPL-MP. Each point corresponds to the winning DAM from a single replication according to ELPL-MP. A replication consisted of (1) simulating item response data with a fixed guessing, a random mean item easiness, and a random log median item discrimination; (2) fitting a 1PL, 2PL, 3PL DAM to that data, and (3) determining the best DAM according to ELPL-MP.

vector $[\mu_{\theta_1} = 0, \mu_{\theta_2} = 0]$ and covariance matrix $\begin{bmatrix} 1 & v \\ v & 1 \end{bmatrix}$. Accordingly, $v$ 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 we might expect $v$ to be high. We can think of $v$ 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. We drew $v$ from a continuous uniform distribution, $v \sim \text{unif}(0,1)$.

## Results for Simulation Study 4

Figure 5 shows the winning DAM 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. We focus here on the role of the correlation between factors, $v$. In general, as $v$ increased, the 1F 2PL was more likely to win. As with Simulation Study 2, we can read these results 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's 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.

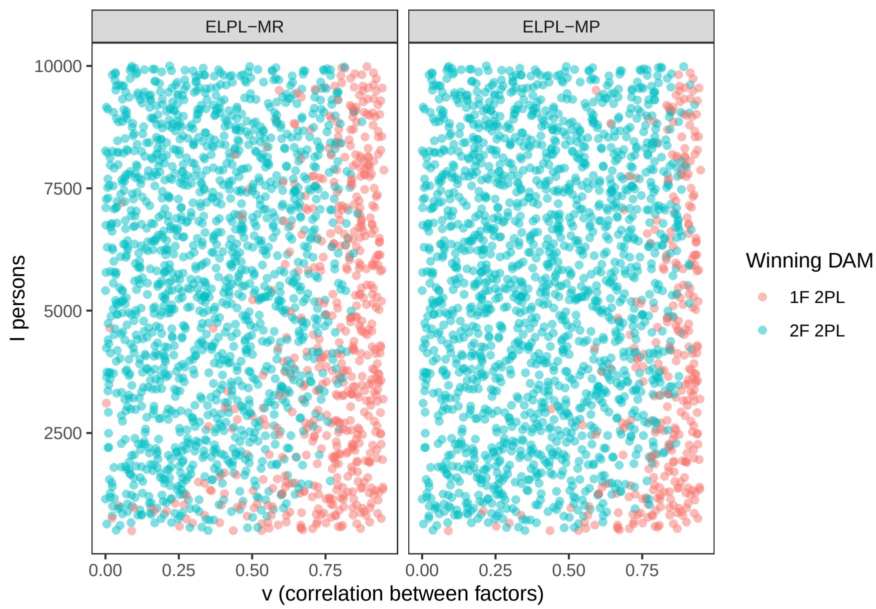


*Figure 5.* Simulation Study 4 results. Each point corresponds to the winning DAM from one of 2000 runs according to each of ELPL-MP (left) and ELPL-MR (right). A replication consisted of (1) simulating item response data with a random number of persons and a random correlation between two factors; (2) fitting a 1F 2PL and 2F 2PL DAM to that data; and (3) determining the best DAM according to ELPL-MP and ELPL-MR.

## 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 DAM could have been the DGM (e.g., whether the expected contingency table from the DAM is similar to a contingency table of the data). We advocated for an alternative view of fit, predictive fit, based on how well a DAM 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 artificial case in which the DGM is a known item response model as is often the case in item response simulation studies. As we describe below, we believe that these predictive fit metrics are useful for evaluating item response models in simulation studies; that our results offer guidance with regard to minimum sample size requirements for item response models; and that predictive fit metrics can help lay the groundwork for future advances in item response model evaluation in practice.

How should DAMs be evaluated and compared in simulation studies? For example, Kang and Cohen (2007) fit both a 2PL and 3PL DAM to data from a 3PL DGM. How should they have decided whether the 2PL DAM or the 3PL DAM fit better? They assumed that because the 3PL DGM was used that the 3PL DAM must fit better. Based on this assumption, they, for example, warned against using a model selection method, BIC, in the conditions in which it frequently selected the 2PL DAM. An alternative is to consider predictive fit by determining which DAM 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 DAM frequently actually makes better predictions than the 3PL DAM, and therefore has better predictive fit. Thus, it is a feature, not a bug, for BIC to select the 2PL DAM 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, Lissak, and Drasgow (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" DAM fit to item response data with fewer than 1000 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 model, which is precisely what our first three simulation studies investigated. 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, defining out-of-sample according to what we have called "missing responses", 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. 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.

Perhaps most importantly, we believe that predictive fit metrics can play a valuable role in laying the groundwork for future advances in item response model evaluation in practice. Psychometricians typically compare item response models using information criterion (e.g., AIC and BIC). These methods, which are based on marginalized likelihoods[12] (Maydeu-Olivares, 2013), can be viewed as approximating ELPL-MP (Stone, 1977). McDonald and Mok (1995) warned that AIC and BIC may fail with modest sample sizes or misspecified models. Cross-validation, which has fewer assumptions, may be better in these cases. The essential logic of cross-validation is that the empirical data is split into a training (in-sample) data and a testing (out-of-sample) data (Stone,

[12] i.e., Ability is treated as a nuisance variable and is integrated over when calculating likelihood.

1974). The models are estimated using the training data and their performance is evaluated by how well they predict the testing data. Bolt and Lall (2003) introduced a marginalized version of cross-validation for item response models where half of the *persons* are randomly assigned to the training data and the other half are assigned to the testing data. The training data is used to estimate item parameters, and the model fit is evaluated according to the marginalized out-of-sample likelihood of the testing data. This method, which we call marginalized cross-validation, can be viewed as potentially providing a better estimate of ELPL-MP than information criterion.[13] Researchers from other fields tend to cross-validate item response models by randomly assigning *item responses* to the training or testing data (Bergner et al., 2012; Wu, Davis, Domingue, Piech, & Goodman, 2020). This version of cross-validation can be viewed as providing an estimate of ELPL-MR. In general, it seems to be the case that psychometricians tend to evaluate item response models using methods that estimate ELPL-MP whereas researchers from other fields use methods that estimate ELPL-MR. Our simulation study results show that ELPL-MP preferences more flexible models, which suggests that psychometricians may be more likely to choose more complicated item response models. Regardless of field, more research is needed to guide IRT practitioners in using cross-validation. For example, answers to the following questions will be useful: How many folds are necessary in k-fold cross-validation? How much better do estimates of ELPL get as more folds are used? Is leave-one-item response-out cross-validation worth the computational expense?

I made this and below better

We close with a fundamental question: How should item response models be evaluated and compared in practice? Should information criterion, cross-validation where the empirical data is split at the *person level*, or cross-validation where the data is split at the *item response level* be used? We believe that the answer must depend 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

[13] We view the conditions under and the degree to which this is true as an open research question.

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 isn't the only consideration (Vehtari, Gelman, & Gabry, 2017). In our view, looking for a connection between predictive fit metrics and practical item response model tasks is, perhaps, the most promising direction for future research. For instance, we hypothesize that ELPL-MR may be a better predictive fit metric if the goal has to do with person abilities, as is typically the case with item response models (Lord, 1986). Demonstrating a link between the two could be hugely valuable because, in practice, estimating person ability error is difficult, if not impossible, whereas estimating the predictive fit metrics is relatively straightforward using methods like cross-validation.

## References

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

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

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.

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

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

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.

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

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

Lord, F. M. (1986). Maximum likelihood and bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*(2), 157–162.

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.

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

Maydeu-Olivares, A., & 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.

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

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

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

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 221–242.

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

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

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 111–133.

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

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.

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

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

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.

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

Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'*. Retrieved from 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*.