Proceedings of the 89th Annual International Meeting of the Psychometric Society, Prague, Czech Republic (2024).

Item response function variability: A strategy for model comparison research in IRT

X. Chen and L. Feuerstahler*

Department of Psychology, Fordham University, Bronx, New York, USA
*Corresponding author. Email: lfeuerstahler@fordham.edu

Abstract

Model complexity is defined as the ability of a model to fit various data patterns. The influence of model complexity on item response theory (IRT) models has historically been explored through simulation studies. However, simulation studies that generate items from different IRT models may confound the inherent differences in the models with artificial differences induced by the choice of data-generating model parameters. In this paper, we introduce the concept of item response function (IRF) variability which can be leveraged to make items from different IRT models as similar as possible in simulation research. Specifically, we illustrate how the distribution of IRF maximum slopes and locations can be harmonized across models. We illustrate this concept with three unidimensional models: the two-parameter model (2PL) model, the negative log-log (NLL) model, and the logistic positive exponent (LPE) model. Illustrative results are presented, followed by an overall discussion.

Keywords: item response theory, goodness-of-fit, model complexity, model comparison, simulation studies

Simulation studies are widely used in item response theory (IRT) research to explore the behavior of models and estimation techniques under known data-generating conditions. When comparing different IRT models in simulation research, it is important to consider the complexity of each model. Model complexity, according to Myung (2000), represents a model's ability to fit different data patterns and is related to two components, the number of parameters and the functional form. In IRT, these two components distinguish different item response functions (IRFs), and therefore different IRT models differ in their ability to provide good model-data fit (Bonifay & Cai, 2017). Previous research considers model complexity in terms of model fit (Bonifay & Cai, 2017; Preacher, 2006). In simulation studies, however, model complexity may also have an impact on the data-generating process because models with different complexity will generate different data patterns. In this paper, we explore the influence of model complexity on IRT simulation results when data are generated from different models. Specifically, we propose the concept of item response function (IRF) variability as a way to quantify and control the differences between sets of IRFs. We illustrate how IRF variability may be controlled across data-generating models in simulation research and explore the impact of controlling IRF variability on simulation results.

1. Item Response Model Complexity

Model complexity, as defined by Myung et al. (2005) and Myung (2000), is the inherent flexibility of a model, or in other words, the ability of a model to describe a variety of data patterns in the complete data space. Model complexity is also called fitting propensity (Preacher, 2006), and is affected by both parameter counts and the functional form (Myung, 2000). Previous research investigated the influence of model complexity on model goodness of fit (GoF; Myung et al., 2005; Myung, 2000;

Pitt and Myung, 2002) and found that complex models often had problems with overfitting and lack of generalizability, even when they performed well in terms of GoF. In other words, overfitting implies that the model learned too much from the noise in the observed data, thereby making it difficult to predict unobserved cases or to generalize to other conditions. As such, researchers typically compare models using information criteria that aim to penalize GoF by model complexity. That is, they attempt to balance model fit and parsimony, in the hope of achieving results that are more generalizable to future data. However, commonly used information criteria such as Akaike information criterion (AIC; Akaike, 1998) and Bayesian information criterion (BIC; Schwarz, 1978) penalize GoF by a function of the number of parameters but do not penalize based on the inherent flexibility of the functional form. Other methods, like minimum description length (MDL; Rissanen, 1987) and information-theoretic measure of complexity (ICOMP; Bozdogan, 1990), consider both parameter counts and functional form. However, to the best of our knowledge, the MDL and ICOMP methods have not been tested for IRT models and are not available in popular IRT software packages such as `mirt` (Chalmers, 2012). As such, the analyses in this paper use two widely-used model comparison indices, AIC and BIC. Previous research (Myung, 2000; Pitt & Myung, 2002) has shown it is essential to consider model complexity when selecting the appropriate model for a particular data set. It is also worth considering model complexity, including parameter counts and function form, when generating data patterns in simulation analysis.

2. IRF Variability

To more fully understand functional form-related model complexity in IRT, we can think of the variety of IRFs that are generated from different models and different distributions of model parameters. For item i, define a linear function of the latent trait θ, $Z_i(\theta) = \alpha_i(\theta - \beta_i)$ where $\alpha_i > 0$. For many dichotomous IRFs, the endorsement probability for item i can be expressed as follows:

$$P(Y_i = 1|\theta) = f(Z_i(\theta)), \tag{1}$$

where Y is the observed item response and f is a monotonically increasing link function. Different IRT models can be defined by different choices of the link function f. For example, the two-parameter logistic (2PL; Birnbaum, 1968) model sets f equal to the logistic function, and the three-parameter logistic (3PL; Birnbaum, 1968) model and the logistic positive exponent (LPE; Samejima, 2000) models modify the 2PL by introducing a third parameter: the third parameter of the 3PL model controls the lower asymptote of the IRF and the third parameter of the LPE model controls the asymmetry of the IRF. In this framework, different choices of f will lead to IRFs with different curves for the same value of Z_i. For example, several authors have noted that items with the same α_i parameters but applied to different models will differ in their slopes (De Ayala, 2013; Molenaar, 2015). In this paper, we quantify IRF similarity by comparing features of the IRF curve P rather than the similarity of item parameters α_i and β_i. In particular, we propose the concept of IRF variability, which considers the distributions of IRF features that can be compared across models. In this paper, we will consider the distributions of the maximum IRF slope and the θ value (location) at which the maximum IRF slope is observed.

If f is continuous and three-times differentiable, the point of maximum slope can be found by setting $\frac{\partial^2 P_i(\theta)}{\partial\theta^2} = 0$, solving that equation for θ, and ensuring that $\frac{\partial^3 P_i(\theta)}{\partial\theta^3} < 0$ for that θ value. For models that follow Equation 1, the IRF slope at θ equals

$$\frac{\partial P_i(\theta)}{\partial\theta} = \alpha_i f(Z_i)(1 - f(Z_i)). \tag{2}$$

Next, note that the point of maximum slope must occur at a stationary point of the first derivative.

Stationary points of $\frac{\partial P_i(\theta)}{\partial \theta}$ may be found by setting

$$\frac{\partial^2 P_i(\theta)}{\partial \theta^2} = \alpha_i^2 f(Z_i)(1 - f(Z_i))(1 - 2f(Z_i)) = 0 \tag{3}$$

and solving for θ. Table 1 displays the functional form f, along with the θ of maximum slope and the value of the maximum slope for the three IRT models considered in this study: the 2PL, LPE, and negative log-log (NLL; Shim et al., 2024). Derivations of these values are provided in Appendix 1. Note that for the 2PL model, the point of maximum slope occurs where $f(Z_i) = 1/2$, but this is not the case for all models. For the NLL model, the slope is maximum when $P = \exp(-\exp(0)) = .368$, and for the LPE, this point depends on all three of its item parameters.

Table 1. Maximum Slopes and Corresponding Locations for Three IRT Models

Model	$f(Z_i)$	θ of maximum slope	maximum slope
2PL	$\text{logit}^{-1}(Z_i)$	$\theta = \beta_i$	$.25\alpha_i$
LPE	$\text{logit}^{-1}(Z_i)^{\xi_i}$	$\theta = \beta_i + \ln(\xi_i)/\alpha_i$	$\alpha_i(\frac{\xi_i}{1+\xi_i})^{1+\xi_i}$
NLL[a]	$\exp(-\exp(-Z_i))$	$\theta = \beta_i$	$\exp(-1)\alpha_i$

a The NLL has previously (Shim et al., 2024) been described as a one-parameter model where all $\alpha_i = 1$. Instead, we use the "two-parameter" version throughout.

To illustrate the differences in distributions of maximum slopes and corresponding locations for different models, we simulated 1,000 items from each of the three models that are listed in Table 1. Population parameters for these models were all simulated from $\alpha \sim LN(0, 0.5)$ and $\beta \sim N(0, 1)$, and for the LPE model, $\xi \sim LN(0, 0.5)$. Density plots of these results are illustrated in Figure 1.

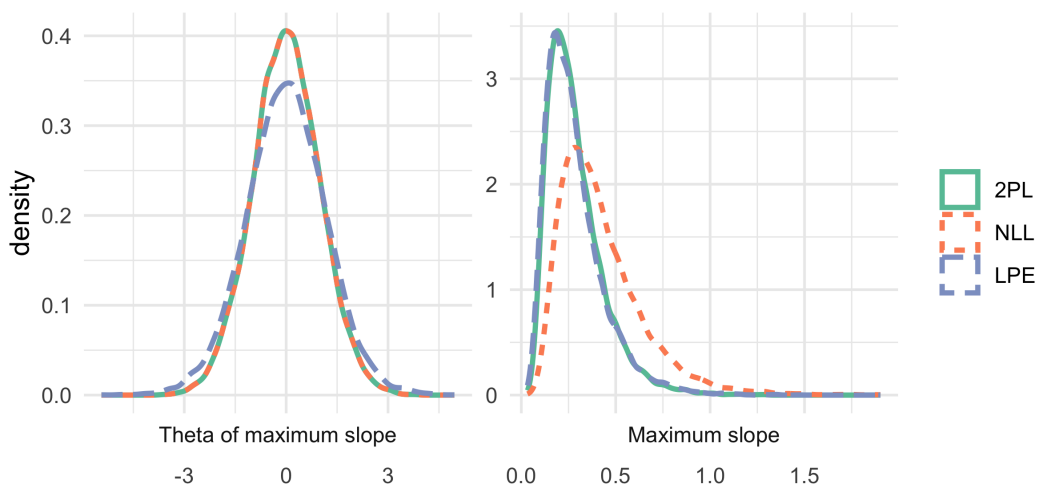

Figure 1. Distributions of highest slopes and corresponding locations

The right panel of Figure 1 illustrates the analytic results from Table 1. The maximum slopes of the 2PL model and the NLL model are determined entirely by α_i but to different extents. For the same α_i value, the slope of the 2PL model is always lower than that of the NLL model. This inference

is consistent with the illustration that the distribution of the NLL maximum slopes is right-shifted compared with the distribution of the 2PL maximum slopes. For the LPE model, the maximum slopes are determined by both α_i and ξ_i. Recalling that the 2PL is a special case of the LPE when $\xi_i = 1$, we find that when $\xi_i < 1$, the maximum slope is lower than that of the 2PL and when $\xi_i > 1$, it is larger.

The left panel of Figure 1 shows the distribution of θ of maximum slopes and shows results consistent with Table 1. Specifically, the corresponding θs of maximum slopes are the same across the 2PL model and the NLL model since these quantities are determined identically in the two models. For the LPE model, the θ of the maximum slope is determined by all three item parameters, leading to a somewhat different distribution of locations.

In this paper, we explore how item parameter data-generating distributions may affect IRT simulation results, especially when data are generated from multiple population models. The 2PL model and the LPE model have similar link functions but different parameter counts, and the 2PL model and the NLL model have the same parameter counts but different link functions. Therefore, to investigate the influence of different function forms, we will compare the 2PL model and the NLL model in Study 1; to investigate the influence of different parameter counts, we will compare the 2PL model and the LPE model in Study 2. More specifically, we will investigate whether controlling the IRF maximum slope and location will better control for differences among the compared models.

3. Simulation Studies

Simulation studies have been widely used to compare IRT models in various ways. Some simulation research has used one model to simulate data and fit the simulated data to one or more candidate models (Lee & Bolt, 2018a, 2018b; Molenaar, 2015). These designs are appropriate to test the accuracy of software or estimation methods (one fitted model) or to investigate the extent to which multiple fitted models can recover features of the data-generating model. Another simulation design generates data according to two or more models and fits those data to one or more models (e.g., Fujimoto and Falk, 2024; Kang and Cohen, 2007; Zhang et al., 2022). This design is typically used to compare the relative accuracy or efficiency of different models, and is often used in model selection research. The methods proposed in this study are relevant to simulation studies that follow this second design.

To our knowledge, prior simulation research in IRT has not considered controlling IRF variation when comparing different models fitted to the same data. In the remainder of this paper, we present two simulation studies that explore the effects of harmonizing the distributions of slopes and locations of IRFs across different data-generating models. The first study compares the 2PL to the NLL. Because these two models include the same number of item parameters, this comparison will allow us to evaluate the effect of controlling parameter distributions with the number of parameters held constant. In this study, we generated parameters based on four model conditions: (a) 2PL,(b) NLL, (c) adjusted NLL that matches the slope and location properties of (a), and (d) adjusted 2PL that matches the slope and location properties of (b). In this way, the IRFs generated by the adjusted 2PL model should be more similar to the IRFs generated by the NLL model than those from the unadjusted 2PL model. The second study compares the 2PL, which has two parameters per item, to the LPE, which has three parameters per item. Here, because the 2PL is a special case of the LPE, we consider only three model conditions: (a) 2PL, (b) LPE, and (c) adjusted 2PL that matches the slope and location properties of the LPE. Details of the data-generating process and adjustments are described in their respective subsections.

For each data-generating model and its corresponding item parameter distributions, we generated 1000 sets of 10 items. For Study 1, item response data were simulated for all four conditions with both $N = 500$ and $N = 1000$ examinees and $\theta \sim N(0, 1)$. Then, each set of item response data was fit to both the 2PL and NLL models using marginal maximum likelihood estimation as implemented in

the `mirt` package. Note that the NLL is not currently available in the `mirt` packages, and so we used its `createItem` functionality to implement this model. Code to implement the NLL model in `mirt` is provided in Appendix 2. In Study 2, because the LPE is known to be difficult to fit (Lee & Bolt, 2018a), item response data were simulated for all three conditions with $N = 10,000$ and $N = 50,000$ examinees and $\theta \sim N(0, 1)$. Then, each set of item response data was fit to both the 2PL and LPE models using the `mirt` package. Marginal maximum likelihood estimation was used to fit the 2PL models, but a $N(0, .5)$ prior was applied to the $\ln(\xi)$ parameter for the LPE model to improve model convergence.

We compared the fitted models in terms of item response function accuracy and GoF. To evaluate recovery, we compared the true and estimated IRFs by the root integrated mean square error (RIMSE; Ramsay, 1991), which is defined as:

$$RIMSE = \sqrt{\int (P_1(Y_i = 1|\theta) - P_2(Y_i = 1|\theta))^2 g(\theta) d\theta}.$$

where P_1 and P_2 are the population and estimated IRFs to compare in the analysis, and $g(\theta)$ represents the θ distribution to integrate over. We calculated the RIMSE separately for each fitted model and compared the distribution of RIMSE values across different data-generating and fitted model conditions. For the analyses presented in this paper, we used a standard normal $g(\theta)$. Smaller RIMSE value indicates the two IRFs are closer to each other (Feuerstahler, 2021). To evaluate penalized goodness-of-fit, we used the AIC and BIC because they penalize goodness-of-fit by parameter counts.

It is important to note that this method of comparing different IRT models assumes that all fitted models are on the same latent trait scale. However, it is often unlikely that this will strictly be true. For example, Shim et al., 2024 noted that the one-parameter NLL is a nonlinear transformation of the one-parameter logistic model. In other words, these two models are theoretically capable of making the exact same predictions as each other, corresponding to different underlying latent trait distributions. However, in practice, these models are often identified by specifying a standard normal prior for θ during model fitting, hindering the ability of the different models to be exact transformations of each other. Moreover, much IRT model comparison research, including that presented in Shim et al., 2024, does not explicitly take this potential confounding factor into account, and it is unclear how large the effects of scale differences may be. As such, we acknowledge that differences in scale are a potential confounding factor that is not accounted for in the following studies, but we still find value in controlling for the IRF features that we do control for.

3.1 Study 1

In this study, we compared parameter recovery and GoF for the 2PL and NLL. To do so, we generated data under four conditions. First, in condition (a) we generated 2PL data with $\alpha \sim LN(0, 0.5)$ and $\beta \sim N(0, 1)$. In condition (b), we generated NLL data with the same data-generating parameters and condition (a): $\alpha \sim LN(0, 0.5)$ and $\beta \sim N(0, 1)$. For the third condition, we aimed to find distributions of NLL parameters that matched the slope and location distributions of the 2PL in condition (a). From Table 1, we know that the θ values at which slopes are maximum equal β_i for both models and so the same distribution can be used to generate β parameters. To align the distributions of maximum slopes, note that the maximum slopes of the 2PL and NLL models are $0.25\alpha_i$ and $\alpha_i \exp(-1)$ respectively. It can be shown (Limpert et al., 2001) that if a variable α is log-normally distributed, $\alpha \sim LN(\mu, \sigma)$, then a constant k times α is also log-normally distributed, $k\alpha \sim LN(\mu + \ln(k), \sigma)$. Therefore, in condition (a), the distribution of maximum slopes is $LN(\ln(0.25), \sigma)$. For the NLL to have the same distribution of maximum slopes, then the NLL $\alpha \sim LN(1 + \ln(0.25), 0.5)$. For the fourth condition, we use the same logic to find a distribution for 2PL α that has the same slope

distribution as condition (b). In condition (b), the distribution of maximum slopes is $LN(-1, 0.5)$, so the 2PL $\alpha \sim LN(-1 - \ln(0.25), 0.5)$. In summary, in condition (c) we generated adjusted NLL (aNLL) parameters from $\alpha \sim (LN(1 + \ln(0.25), 0.5)$ and $\beta \sim N(0, 1)$, and in condition (d) we generated adjusted 2PL (a2PL) parameters from $\alpha \sim LN(-1 - \ln(0.25), 0.5)$ and $\beta \sim N(0, 1)$. In this way, conditions (a) and (c) will have systematically lower maximum slopes than conditions (b) and (d).

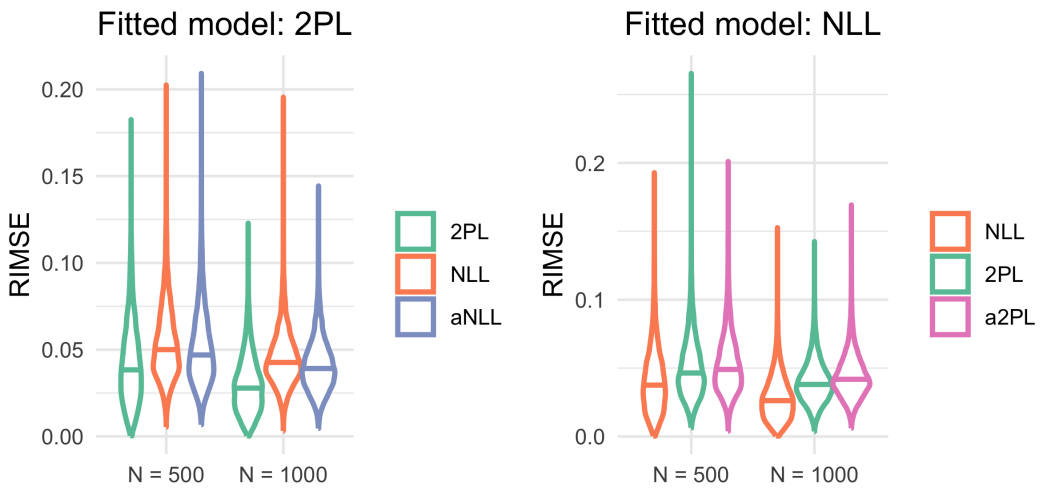

Figure 2. Violin plots of RIMSEs for the 2PL, NLL, aNLL, and a2PL. The central lines of each violin plot indicate medians. Different colors are used for different data-generating models.

The distributions of RIMSEs are shown in Figure 2. The left plot shows results from data fit to the 2PL and the right plot shows results from data fit to the NLL. As may be expected, the lowest RIMSEs are found when the data are fit to the same model that generated the data. However, these plots illustrate that different sets of data-generating parameters can yield different sets of results for the same data-generating model. To be more specific, consider the left plot with $N = 1000$ where data is fit to the 2PL. The mean RIMSE equals .030 for the 2PL, .045 for the NLL, but only .041 for the aNLL. Although the difference between RIMSEs for the NLL and aNLL is small, it is statistically significant, $t(19987) = 18.08, p < .001$. When data are generated according to the aNLL, the difference in IRF recovery is not as large as with the NLL, suggesting that the difference between the two models might be exaggerated if the adjustment is not used. Similarly, consider the right plot with $N = 1000$ where data is fit to the NLL. The mean RIMSE equals .029 for the NLL, .040 for the 2PL, and .044 for the a2PL. Again, the difference between RIMSEs for the 2PL and a2PL is statistically significant, $t(19982) = 18.54, p < .001$. The same small but statistically significant differences are also observed for the $N = 500$ sample size.

It is notable that when data were fit to the 2PL, the adjusted model led to RIMSE results that are more similar to those of the data-generating model. However, when data were fit to the NLL, the adjusted model led to RIMSE results that are *less* similar to those of the data-generating model. Not only this but the magnitude of the difference in average RIMSE for the 2PL and NLL differs across the two sets of adjustments. Although not the main focus of this paper, we suspect that this difference occurs because of the difference in average slopes for the 2PL and aNLL versus the NLL and a2PL.

We next consider the ability of information criteria to select the correct model when comparing the 2PL and NLL models. Note that, as implemented in this paper, the 2PL and NLL have the same number of parameters such that the same penalty term will be imposed for both models. For this

reason, both the AIC and BIC will always select the model with the highest log-likelihood and will always give the same model selection results as each other. Therefore, the model selected by AIC or BIC must be that with the highest log-likelihood. Due to its prevalence in model selection, we chose to present these results in terms of the AIC. The percentages of replications for which AIC selected the data-generating model are shown in Table 2.

Table 2. Percentage of Replications for which AIC Selected the Data-Generating Model

	2PL	a2PL	χ^{2a}	NLL	aNLL	χ^{2a}
N = 500	81.2	90.5	34.84*	92.9	88.8	9.62*
N = 1000	91.5	97.1	28.14*	98.1	94.5	17.19*

a χ^2 tests of equal proportions with continuity correction with 1 degree of freedom. $\star$ indicates $p < .05$.
b "2PL", "a2PL", "NLL", and "aNLL" indicate the data-generating model.

In all cases, the data-generating model was selected significantly more often when the data was generated according to models with higher slopes. This finding coincides with previous research (Lopez Rivas et al., 2009) demonstrating that high-discrimination items tend to make better anchor items for differential item functioning analysis, possibly because these items provide more information and lead to better model identification. These results suggest that if the slopes are not adjusted for in the data-generating process, the differences between models may look larger than they actually are. For example, if N = 1000 and data are generated according to conditions (a) and (b) (i.e., the 2PL and NLL with the same distributions of α and β), it appears the NLL is more capable of selecting the correct model (98.1%) than the 2PL (91.5%). However, if the NLL is compared to the a2PL instead (condition d), then the 2PL is correctly selected in 97.1% of replications, a difference that is not statistically significant from that for the NLL, $\chi^2(1) = 1.73, p = .19$. Alternatively, if the 2PL is compared to the aNLL, then the aNLL is correctly selected in 94.5% of cases, which when compared to the 2PL, reflects a statistically significant difference, $\chi^2(1) = 6.46, p = .01$, though a smaller difference than that found between the 2PL and (unadjusted) NLL.

3.2 Study 2

In the second study, we compared parameter recovery and penalized GoF for the 2PL and LPE models. To do so, we generated data under three conditions. As in Study 1, in condition (a) we generated 2PL data with $\alpha \sim LN(0, 0.5)$ and $\beta \sim N(0, 1)$, and in condition (b), we generated LPE data from the same distributions of α and β and from $\xi \sim LN(0, 0.5)$. In our condition (c), we found adjusted 2PL (a2PL) parameters that match the distributions of maximum slopes and locations from condition (b). Note that we did not attempt to find an adjusted LPE model that matched the maximum slopes from condition (a) because the 2PL is a special case of the LPE, and such matching would degenerate to the 2PL. To find the parameters for the a2PL, there is now no convenient distribution to sample from, but we can derive appropriate parameters based on the parameters drawn in condition (b). To do so, we first equate the maximum slopes to find α_i for the a2PL. Here,

$$\alpha_{i,\text{a2PL}} = 4\alpha_{i,\text{LPE}}\left(\frac{\xi_i}{1+\xi_i}\right)^{1+\xi_i}. \tag{4}$$

We can next find the β_i parameters for the a2PL can be found by equating the locations of the maximum slope:

$$\beta_{i,\text{a2PL}} = \beta_{i,\text{LPE}} + \ln(\xi_i)/\alpha_{i,\text{LPE}}. \tag{5}$$

As such, a2PL parameters were found by first generating LPE parameters as in condition (b), and then using Equations 4 and 5 to find 2PL parameters. From Figure 1, we know that the distributions

of maximum slopes are already quite similar for the 2PL and LPE without any adjustment, but that the θ of maximum slope is more variable for the LPE (and thereby for the a2PL).

Figure 3 displays violin plots for datasets fit to the LPE (left) and to the 2PL (right). The most striking result of this comparison is that the RIMSE distributions for data fit to the LPE are highly skewed, a consequence of the known difficulties associated with fitting this model resulting from the high within-item correlation among its parameters (Liao & Bolt, 2021). As a result, we focus this discussion on medians rather than means. When data are fit to the LPE, we find that the lowest median RIMSE is observed for data generated from the LPE model. This is despite the fact that, because the 2PL is a special case of the LPE, the LPE should be able to describe the response patterns from the 2PL and a2PL data just as accurately as the LPE data. In addition, according to Mann-Whitney tests, we do not find significant differences in median RIMSEs for data generated from the 2PL versus a2PL in any sample size or fitted model condition.

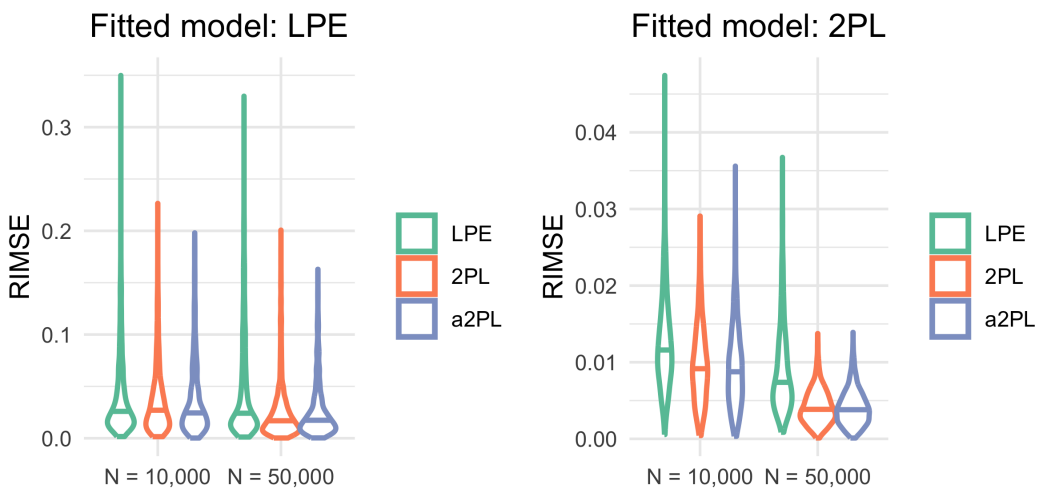

Figure 3. Violin plots of RIMSEs for the LPE, 2PL, and a2PL. The central lines of each violin plot indicate medians. Different colors are used for different data-generating models.

For nearly all fitted models in Study 2, both AIC and BIC selected the 2PL. The only exception to this pattern occurred when data were generated from and fit to the LPE with $N = 50,000$, for which the LPE was correctly selected by AIC in 4.3% of replications. Because these information criteria nearly always select the 2PL in these comparisons, we instead considered how harmonizing the distribution of slopes and locations affected overall GoF through the log-likelihood itself. Cumulative distributions of log-likelihoods across 1,000 replications are shown in Figure 4. As must be the case since the 2PL is a special case of the LPE, the log-likelihood is greater for the LPE than for the 2PL in all conditions. However, in all cases, the a2PL led to a cumulative distribution of log-likelihoods that was closer to the LPE than was the 2PL. This result suggests that there exist some nonessential differences in fitting propensity between the LPE and 2PL generated from the same distribution of α_i and β_i parameters. By generating data from the a2PL instead of the 2PL, the 2PL and LPE models might be more fairly compared.

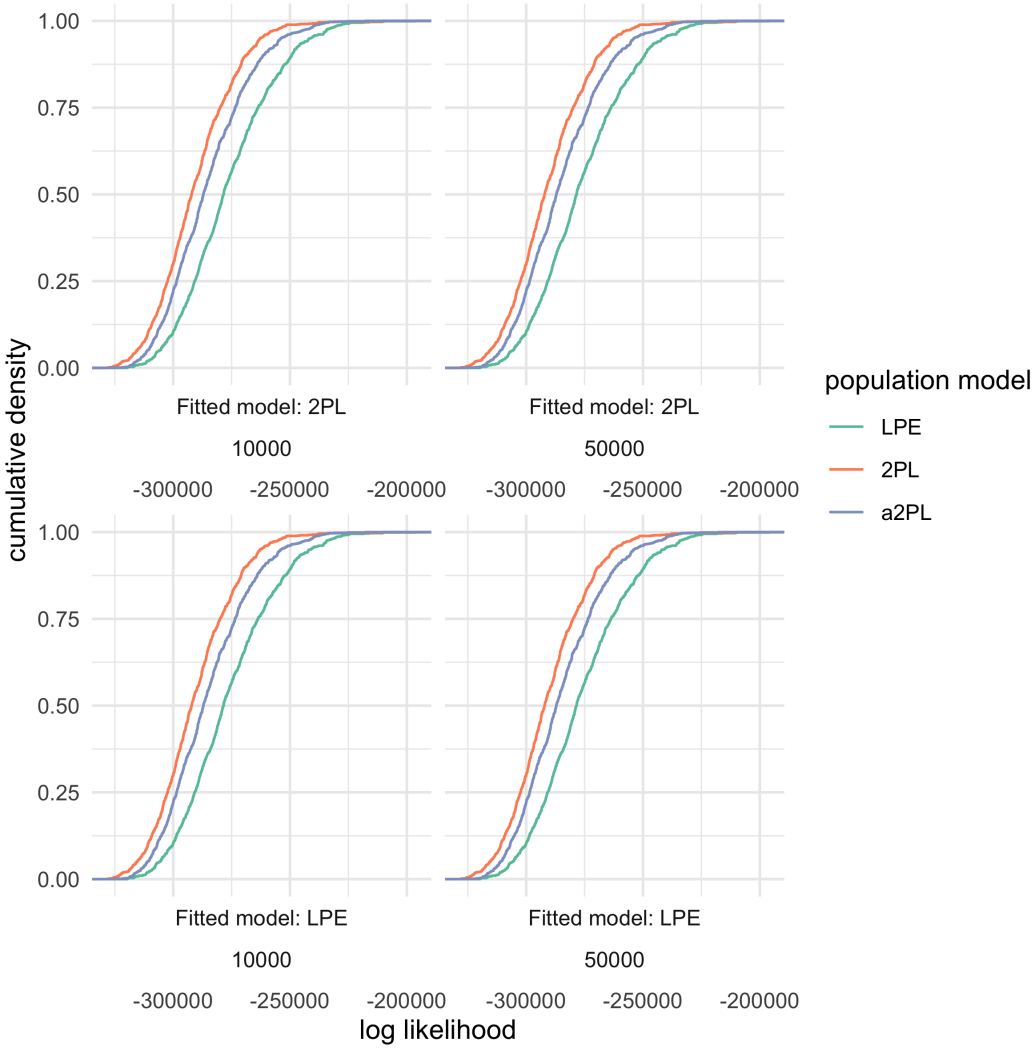

Figure 4. Cumulative distributions of fitted log-likelihoods for the LPE, 2PL, and a2PL.

4. Discussion

In this paper, we did not intend to draw generalizable conclusions about particular models. Instead, we aimed to investigate the implications of controlling the features of IRFs in simulation studies as an alternative to using the same distributions of item parameters for each model. Because item parameters serve different roles in different item response models, we believe that matching the common features of IRFs (here, the maximum IRF slope and its location) generated from different models is a more justifiable way to compare models. In Study 1, we showed that after controlling the maximum IRF slopes and locations, the 2PL model recovered item response functions just as accurately for data generated from the NLL as it did for 2PL-generated data. In addition, penalized goodness-of-fit indices had comparable percentages of selecting the correct data-generating model when controlling for these features, but different rates before controlling for these features. These results suggest that the assumed latent trait distribution does not have a significant influence on the model fit and that misleading conclusions might be reached if these features are not controlled

in simulation studies. The justification for this procedure is closely related to the idea of model complexity. Previous research has shown that model complexity is affected by both the number of parameters and the functional form of a model. The procedure we outline in this paper allows users to more precisely compare the features of model fitting that are associated with differences in functional form.

In this paper, we chose to study the location of maximum slope and the maximum slope itself as two IRF features that are common to most unidimensional models for dichotomous data. However, these features do not apply to all such models and may not be the best choices of common features for polytomous or multidimensional IRT models. For example, IRFs based on the generalized log-log link function (Zhang et al., 2022) might have multiple points of (locally) maximum slope. In addition, the location of an IRF might be better characterized by the point at which $P = .5$ or other criteria.

To extend the idea of IRF variability to polytomous models, consider two commonly used models for ordered polytomous data, the generalized partial credit model (GPCM; Muraki, 1992) and the graded response model (GRM; Samejima, 1969). Because these models represent multiple category response functions per item, there are many potential criteria that could be used to harmonize the features of models. These features include the locations and values of maximum slope for functions that define the boundaries between adjacent categories, points at which the probability of a category response equals the probability of the next highest category, or the features of the functions that define a response probability in an individual category. More research is needed to derive these features for polytomous model and determine the most affective adjustment criteria.

Multidimensional IRT models pose another challenge. Unlike univariate IRT models for which IRFs can be illustrated with two-dimensional plots, as the number of dimensions increases, it is more and more difficult to illustrate the IRFs. Like polytomous models, multidimensional models will pose a challenge in that there are multiple features per item that a researcher might want to align. A possible starting point is to use multidimensional difficulty and discrimination (Reckase & McKinley, 1991) as a feature common across multidimensional models, but more research is needed to further explore this idea.

5. Conclusion

In conclusion, this paper explores how the choice of item parameters from which to generate IRT simulation data may affect simulation results. We illustrated this idea in terms of the distributions of maximum slopes and corresponding θ values for three unidimensional IRT models for dichotomous data. We demonstrated that adjusting the data-generating parameters for a model to have similar feature distributions as another model can affect the results of simulation research comparing different data-generating models.

Competing Interests None

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike* (pp. 199–213). Springer.

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

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate behavioral research*, *52*(4), 465–484.

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics-Theory and Methods*, *19*(1), 221–278.

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

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Feuerstahler, L. (2021). Flexible item response modeling in r with the flexmet package. *Psych*, *3*(3), 447–478.

Fujimoto, K. A., & Falk, C. F. (2024). The accuracy of bayesian model fit indices in selecting among multidimensional item response theory models. *Educational and Psychological Measurement*, *84*(2), 217–244.
Kang, T., & Cohen, A. S. (2007). Irt model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331–358.
Lee, S., & Bolt, D. M. (2018a). An alternative to the 3pl: Using asymmetric item characteristic curves to address guessing effects. *Journal of Educational Measurement*, *55*(1), 90–111.
Lee, S., & Bolt, D. M. (2018b). Asymmetric item characteristic curves and item complexity: Insights from simulation and real data analyses. *Psychometrika*, *83*(2), 453–475.
Liao, X., & Bolt, D. M. (2021). Item characteristic curve asymmetry: A better way to accommodate slips and guesses than a four-parameter model? *Journal of Educational and Behavioral Statistics*, *46*(6), 753–775.
Limpert, E., Stahel, W. A., & Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues: On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: That is the question. *BioScience*, *51*(5), 341–352.
Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement*, *33*(4), 251–265.
Molenaar, D. (2015). Heteroscedastic latent trait models for dichotomous data. *Psychometrika*, *80*(3), 625–644.
Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *Applied psychological measurement*, *16*(2), 159–176.
Myung, I. J., Pitt, M. A., & Kim, W. (2005). Model evaluation, testing and selection. *Handbook of cognition*, 422–436.
Myung, I. J. (2000). The importance of complexity in model selection. *Journal of mathematical psychology*, *44*(1), 190–204.
Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in cognitive sciences*, *6*(10), 421–425.
Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, *41*(3), 227–259.
Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*(4), 611–630.
Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied psychological measurement*, *15*(4), 361–373.
Rissanen, J. (1987). Stochastic complexity and the mdl principle. *Econometric Reviews*, *6*(1), 85–102.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(S1), 1–97.
Samejima, F. (2000). Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves. *Psychometrika*, *65*, 319–335.
Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.
Shim, H., Bonifay, W., & Wiedermann, W. (2024). Parsimonious item response theory modeling with the negative log-log link: The role of inflection point shift. *Behavior Research Methods*, *56*(5), 4385–4402.
Zhang, J., Zhang, Y.-Y., Tao, J., & Chen, M.-H. (2022). Bayesian item response theory models with flexible generalized logit links. *Applied Psychological Measurement*, *46*(5), 382–405.

Appendix 1. Derivations

In this paper, we consider IRT models that take the form,

$$P(Y_i = 1|\theta) = f(Z_i(\theta)),$$

where $Z(\theta) = \alpha(\theta - \beta)$, $\alpha > 0$, and f is a differentiable monotonically increasing function. The slope of the IRF is given by

$$\frac{\partial P}{\partial \theta} = \frac{\partial f(Z(\theta))}{\partial \theta} = \alpha \frac{\partial}{}$$

.

The point of maximum slope occurs at the θ value for which $\frac{\partial^2 f(Z(\theta))}{\partial \theta} = 0$ and $\frac{\partial^3 P}{\partial \theta^3} < 0$. Below, we derive the θ value at which the IRF slope is maximum and the maximum slope for the three models discussed in this paper: the 2PL, the LPE, and the NLL.

Appendix 1.1 2PL

For the 2PL, the IRF and its first, second, and third derivatives are given as follows:

$$P_i(\theta) = \frac{1}{1 + \exp(-Z_i(\theta))},$$

$$\frac{\partial P_i(\theta)}{\partial \theta} = \alpha_i P_i(\theta)(1 - P_i(\theta)), \tag{6}$$

$$\frac{\partial^2 P_i(\theta)}{\partial \theta^2} = \alpha_i^2 P_i(\theta)(1 - P_i(\theta))(1 - 2P_i(\theta)), \tag{7}$$

and

$$\frac{\partial^3 P_i(\theta)}{\partial \theta^3} = \alpha_i^3 P_i(\theta)(1 - P_i(\theta))(1 - 6P_i(\theta) + 6P_i(\theta)^2). \tag{8}$$

Assuming $\alpha_i > 0$, Equation 7 = 0 whenever $P_i(\theta) = 0$, $P_i(\theta) = 1$, or $P_i(\theta) = \frac{1}{2}$. The first two solutions cannot be the point of maximum slope since they are at the asymptotes of the IRF and lead to $\frac{\partial^3 P_i(\theta)}{\partial \theta^3} = 0$. Therefore, the point of maximum slope for the 2PL is at $P_i(\theta) = \frac{1}{2}$, which will occur when $\theta = \beta_i$. Evaluating Equation 8 at this point yields a third derivative equal to $-\frac{\alpha_i^3}{8}$, which is negative whenever $\alpha_i > 0$. Finally, the maximum slope is found by evaluating Equation 6 at $\theta = \beta_i$, which yields a maximum slope of $.25\alpha_i$.

Appendix 1.2 NLL

For the NLL, expressions are simplified by setting $Z_i = Z_i(\theta)$ and $Z_i^\star = \exp(-Z_i)$. The IRF for the NLL and its first, second, and third derivatives are given as follows:

$$P_i(\theta) = \exp(-Z_i^\star),$$

$$\frac{\partial P_i(\theta)}{\partial \theta} = \alpha_i \exp(-Z_i - Z_i^\star), \tag{9}$$

$$\frac{\partial^2 P_i(\theta)}{\partial \theta^2} = \alpha_i^2 (Z_i^\star - 1) \exp(-Z_i - Z_i^\star), \tag{10}$$

and

$$\frac{\partial^3 P_i(\theta)}{\partial \theta^3} = \alpha_i^3 [(Z_i^\star - 2) \exp(-2Z_i - Z_i^\star) + (1 - Z_i^\star) \exp(-Z_i - Z_i^\star)]. \tag{11}$$

Equation 10 = 0 whenever $Z_i^\star = 1$ or when $Z_i^\star = -\infty$. The latter condition will occur when $P_i(\theta) = 0$, which cannot be a point of maximum slope. Therefore, the point of maximum slope will occur at $Z_i^\star = 1$, which is the point at which $Z_i = 0$ and $P = \exp(-1) \approx .37$. Evaluating Equation 11 at this point gives $\frac{\partial P_i(\theta)^3}{\partial^3 \theta} = -\exp(-1)\alpha_i$, which is negative whenever $\alpha_i > 0$. Therefore, the point of maximum slope occurs $Z_i = 0$, that is, when $\theta = \beta_i$. Finally, the maximum slope is found by evaluating Equation 9 at this point, which yields a maximum slope of $\exp(-1)\alpha_i$.

Appendix 1.3 LPE

For the LPE, the IRF equals

$$P_i = P_i(\theta) = (P_i^\star)^{\xi_i}$$

where $P_i^\star$ is IRF for the 2PL. The first, second, and third derivatives of the LPE model equal

$$\frac{\partial P_i}{\partial \theta} = \alpha_i \xi_i P_i (1 - P_i^\star), \tag{12}$$

$$\frac{\partial^2 P_i(\theta)}{\partial \theta^2} = \alpha_i^2 \xi_i P_i(1 - P_i^\star)[\xi_i(1 - P_i^\star) - P_i^\star], \tag{13}$$

and

$$\frac{\partial^3 P_i(\theta)}{\partial \theta^3} = \alpha_i^3 \xi_i P_i(1 - P_i^\star)[(\xi_i(1 - P_i^\star) - P_i^\star)^2 - (1 + \xi_i)P_i^\star(1 - P_i^\star)]. \tag{14}$$

For $\alpha_i > 0$ and $\xi_i > 0$, setting Equation 13 = 0 at $P_i = 0$, $P_i = 1$ or $P_i^\star = \frac{\xi_i}{1+\xi_i}$. Because the first two solutions cannot be the point of maximum slope. Rearranging the third solution, we find that the point of maximum slope occurs when $\xi_i = \exp(\alpha_i(\theta - \beta_i))$, that is, where $\theta = \beta + \ln(\xi)/\alpha$. Plugging this value into Equation 14 yields $\alpha_i^3 P_i(1 - P_i^\star)[-\frac{\xi_i(1+\xi_i)}{(1+\xi_i)^2}]$, which will always be negative for $\xi_i > 0$, $P_i > 0$, and $P_i^\star > 0$. Finally, evaluating Equation 12 at the point of maximum slopes yields a maximum slope of $\alpha_i(\frac{\xi_i}{1+\xi_i})^{1+\xi_i}$.

Appendix 2. mirt Code for the Two-Parameter NLL

```
NLL <- createItem(name = "NLL", par = c(alpha = 1, beta = 0),
                  est = c(TRUE, TRUE),
                  P = function(par, Theta, ncat){
                    alpha <- par[1]
                    beta <- par[2]
                    P1 <- exp(-exp(-(alpha * (Theta - beta))))
                    cbind(1 - P1, P1)
                  })
```