CRESST REPORT 844

SEMI-PARAMETRIC ITEM RESPONSE FUNCTIONS IN THE CONTEXT OF GUESSING

APRIL 2015

Carl F. Falk
Li Cai

National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

Semi-Parametric Item Response Functions in the Context of Guessing

CRESST Report 844

Carl F. Falk
Graduate School of Education & Information Studies
University of California, Los Angeles

Li Cai
CRESST/University of California, Los Angeles

April 2015

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The research reported here was partially supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D140046 to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

To cite from this report, please use the following as your APA reference: Falk, C. F., & Cai, L. (2015). *Semi-parametric item response functions in the context of guessing* (CRESST Report 844). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Semi-parametric item response functions in the context of guessing

Carl F. Falk & Li Cai
University of California, Los Angeles

Semi-parametric item response functions in the context of guessing

Abstract

We present a logistic function of a monotonic polynomial with a lower asymptote, allowing additional flexibility beyond the three-parameter logistic model. We develop a maximum marginal likelihood based approach to estimate the item parameters. The new item response model is demonstrated on math assessment data from a state, and a computationally efficient strategy for choosing the order of the polynomial is demonstrated and tested.

Keywords: filtered polynomial, guessing, item response theory

1 Introduction

In item factor analysis, the three-parameter logistic (3PL; Birnbaum, 1968) model is a commonly used item response model that includes a lower asymptote (top panel in Figure 1). This item model can be particularly useful in educational assessments when the lower asymptote represents the probability of correctly "guessing" the answer to a multiple choice question. But, what happens when there is a non-zero probability of guessing and the item response function (IRF) that generated the data does not neatly follow this functional form (bottom panel in Figure 1)?

It is known that using standard item response models such as the two-parameter logistic (2PL; Birnbaum, 1968) and generalized partial credit models (GPC; Muraki, 1992) when the true IRF follows a different shape can result in poor recovery of the response function, item or model misfit, and suboptimal scoring of individuals' proficiency (e.g., Falk & Cai, in press; Liang, 2007; Liang & Browne, 2015; Ramsay & Abrahamowicz, 1989). We expect these same problems to occur when forcing a 3PL model on item response data that could have been generated by an IRF with a different functional form. Remedies to this problem in general include approaches that allow for more flexibility in the estimated IRF. For example, non-parametric methods may quickly estimate an IRF with good recovery, but with the caveat that the resulting IRF may not be monotonically increasing (Ramsay, 1991). Bayesian non-parametric methods may also be employed (Duncan & MacEachern, 2008, 2013; Miyazaki & Hoshino, 2009; Qin, 1998), but may be slow to estimate (Liang, 2007).

The present research builds upon recent semi- (or quasi-) parametric item response functions that are built by replacing the linear predictor of standard response functions with a monotonic polynomial (Falk & Cai, in press; Liang, 2007; Liang & Browne, 2015). So far, these approaches have been demonstrated to allow more flexibility to the 2PL and GPC item models, and can result in better recovery of IRFs and latent proficiency. Specifically, we will show how a logistic function of a monotonic polynomial with a lower

asymptote (LMPA) can allow for greater flexibility than the standard 3PL. In addition, the item model reduces to the 3PL at the lowest-order polynomial.

An additional issue we address is the selection of the order of the polynomial for each item. Previous research has so far relied mostly on using information criteria (e.g., AIC or BIC) for selecting the order of the polynomial, which requires refitting the model. For example, Falk and Cai (in press) used an AIC step-wise approach where at each iteration, each item in turn was considered as a candidate for being modelled as a higher order polynomial. The item that improved AIC the most was selected as having a higher-order polynomial, and the process then repeats until no progress can be made. When using in conjunction with the EM algorithm (e.g., Bock & Aitkin, 1981), this process can be computationally prohibitive with a long test. Thus, the most items used by Falk and Cai (in press) was 20. Use of a different estimation approach may be faster computationally (Liang, 2007; Liang & Browne, 2015), but has not been demonstrated for cases where multiple group analyses are employed or the dataset contains missing data. As an alternative, we suggest that candidate items may be identified without additional model fitting by considering item fit statistics such as $S - X^2$ (Orlando & Thissen, 2000, 2003). Since $S - X^2$ has been successfully used in previous research to identify atypical IRFs, its use has potential in item screening and may result in refitting the model a fewer number of times. That is, only items that fit poorly according to $S - X^2$ may be good candidates for modeling with higher-order polynomials. Since it is possible that as part of a testing program poor item fit may be used as a flag for deactivating items, this approach also has an advantage if it is able to reduce the number of poorly fitting items.

The remainder of this manuscript is organized as follows. In Section 2, we present the proposed item model, LMPA. Section 3 presents an illustration using a large-scale state math assessment that uses $S - X^2$ to screen items before modeling with the LMPA with higher-order polynomials. Section 4 then presents simulation results illustrating the ability of the LMPA item model to improve IRF recovery in conjunction with $S - X^2$

guided polynomial order selection. Finally, Section 5 contains concluding remarks.

2 The proposed item response model

2.1 Logistic function of a monotonic polynomial with asymptote (LMPA)

To introduce notation, consider $i = 1, 2, \ldots, N$ examinees complete $j = 1, 2, \ldots, n$ dichotomously scored items, with observed item responses $y_{ij} \in [\ 0 \ \ 1\ ]$. One way of writing the 3PL model for the "correct" response is as follows:

$$P(1|\theta_i, \kappa_j, \delta_j, \gamma_j) = c(\kappa_j) + \frac{1 - c(\kappa_j)}{1 + \exp(-(\delta_j + \gamma_j \theta_i))} \tag{1}$$

where δ_j and γ_j are the intercept and slopes, respectively. And the pseudo-guessing parameter $c(\kappa_j)$ determines the lower asymptote, which is an estimate of the proportion of examinees in the latent class adopting the "guessing" strategy, itself reparameterized as a function of κ_j to allow unconstrained estimation:

$$c(\kappa_j) = \frac{1}{1 + \exp(-\kappa_j)}. \tag{2}$$

To allow one or more "bends" in the 3PL model, we can replace the slope with a monotonic polynomial function for item j:

$$P(1|\theta_i, \delta_j, \omega_j, \boldsymbol{\alpha}_j, \boldsymbol{\tau}_j) = c(\kappa_j) + \frac{1 - c(\kappa_j)}{1 + \exp(-(\delta_j + m_j(\theta_i, \omega_j, \boldsymbol{\alpha}_j, \boldsymbol{\tau}_j)))} \tag{3}$$

Here the polynomial (omitting the intercept) is of order $2k + 1$, and its derivative with respect to θ is of order $2k$,

$$m(\theta, \omega, \boldsymbol{\alpha}, \boldsymbol{\tau}) = m(\theta, \mathbf{b}) = b_1\theta + b_2\theta^2 + \cdots + b_{2k+1}\theta^{2k+1} \tag{4}$$

$$m'(\theta, \mathbf{a}) = a_0 + a_1\theta + a_2\theta^2 + \cdots + a_{2k}\theta^{2k} \tag{5}$$

where k is a user-specified non-negative integer, which may vary across items. Thus, if

$k = 0$, the model reduces to the 3PL. To ensure monotonicity of the polynomial (i.e., the function increases as θ increases), the derivative is parameterized such that it is always positive, which entails the coefficients $\mathbf{b} = [\ b_1 \quad \cdots \quad b_{2k+1}\ ]$ are a complicated function of the parameters ω, $\boldsymbol{\alpha}$, and $\boldsymbol{\tau}$. That the polynomial is monotonically increasing leads to a monotonically increasing response function for the correct response. Details of this parameterization are given by Falk and Cai (in press) and are due much to the hard work of others (Liang, 2007; Elphinstone, 1985). In brief, the coefficients $\mathbf{a}_k = [\ a_0 \quad \cdots \quad a_{2k}\ ]$, can be represented in recursive form as:

$$\mathbf{a}_k = \mathbf{T}_k \mathbf{a}_{k-1} = \mathbf{T}_k \mathbf{T}_{k-1} \cdots \mathbf{T}_2 \mathbf{T}_1 \exp(\omega) \tag{6}$$

where the $(2k+1) \times (2k-1)$ matrix $\mathbf{T}_k$ contains only α_k and τ_k:

$$\mathbf{T}_k = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -2\alpha_k & 1 & 0 & \cdots & 0 & 0 & 0 \\ \alpha_k^2 + \exp(\tau_k) & -2\alpha_k & 1 & \cdots & 0 & 0 & 0 \\ 0 & \alpha_k^2 + \exp(\tau_k) & -2\alpha_k & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha_k^2 + \exp(\tau_k) & -2\alpha_k & 1 \\ 0 & 0 & 0 & \cdots & 0 & \alpha_k^2 + \exp(\tau_k) & -2\alpha_k \\ 0 & 0 & 0 & \cdots & 0 & 0 & \alpha_k^2 + \exp(\tau_k) \end{bmatrix}$$

We can then easily obtain the coefficients $\mathbf{b}$ of the polynomial from $\mathbf{a}$ by noticing that $b_u = a_{u-1}/u$ for $u = 1, \ldots, 2k+1$.

2.2 Example response functions

The example response functions in the bottom panel of Figure 1 were based on the LMPA model, with $k = 1$ and $k = 2$, and their item parameters reported in Table 1. These response functions were obtained by analyses conducted on a state math assess-

ment discussed in Section 3. Note that higher-order polynomials tend to result in more flexibility, but could be prone to overfitting noise in the data.

Non-standard response functions such as these can be the result of a number of different reasons. In the context of educational assessments, we propose that data from a heterogeneous population could result in non-standard response functions. Suppose our data come from two different unidentified groups that do not differ in overall proficiency, but differential item functioning exists for a particular item. That is, the item is easier for Group 1 than Group 2 (see Figure 2). Such a case could happen in practice if, for example, the item has unique/specific content that some students were exposed to through instruction, whereas other students were not. The data generating model for the correct response to this item is then a mix of the two group's response functions:

$$P(1|\theta) = p_1 P_1(1|\theta) + p_2 P_2(1|\theta) \tag{7}$$

where p_1 and $p_2 = 1 - p_1$ are the proportion of respondents in each group, and P_1 and P_2 are short-hand for the response functions for Group 1 and 2, respectively. Suppose item parameters of $\kappa = -1.39$ and $\omega = .8$ for both groups, but $\delta = -4$ for Group 1 and $\delta = 4$ for Group 2. If we use mixing proportions of 70% and 30% for Groups 1 and 2, then this results in response functions with non-standard bends (see Figure 2). Although this response function does not come from the LMPA model, the LMPA can still provide a good approximation to it. To illustrate, Figure 3 displays the same mixture IRF (black line) with the best-fitting 3PL line (blue line; left panel) and LMPA line (blue line; right panel) with only a 3rd-order polynomial ($k = 1$). Although the LMPA does not completely overlap with the mixture IRF, the two are almost indistinguishable. Thus, while some approaches do more closely represent the mixing of IRFs or their parameters to achieve flexible IRF shapes (e.g., Duncan & MacEachern, 2008, 2013; Miyazaki & Hoshino, 2009) or explicit identification of latent classes (e.g., Rost, 1990), the LMPA

approach may provide a reasonable alternative solution that can accommodate such shapes. Importantly, it may also serve as a leading indicator followed by more detailed and computationally-demanding analysis.

2.3 Estimation and the use of stabilizing prior distributions

As Falk and Cai (in press) did for other models including the monotonic polynomial, we used the Bock and Aitkin (1981) machinery of maximizing the marginal likelihood using the EM algorithm for use with the LMPA approach. Full first- and second-order derivatives of the LMPA model necessary for estimation are given in Appendix A. Since the LMPA model adds additional parameters to the 3PL, which is known to be difficult to estimate, prior distributions were placed on some model parameters to provide additional stability. Technically, our approach is Bayesian. It constitutes finding the mode of the posterior distribution for model parameters, though we note this approach is common in estimation of the 3PL (Bock, Gibbons, & Muraki, 1988; Cai, Yang, & Hansen, 2011; Mislevy, 1986). It may be best to understand the effects of these priors as stabilizing soft constraints on gradients and ridging/conditioning constants on the Hessian, without completely abandoning the operational efficiencies of maximum marginal likelihood.

Specifically, we used diffuse normal priors on all α and τ parameters as also done by Falk and Cai (in press), such as $\pi(\alpha) \sim \mathcal{N}(0, 50)$ and $\pi(\tau) \sim \mathcal{N}(-1, 50)$. To provide stability to the lower asymptote, we use a prior such as $\pi(\kappa) \sim \mathcal{N}(-1.39, .25)$, where $c(-1.39) \approx .20$, or the rate of guessing we might expect from a five option multiple choice question (see also Cai et al., 2011). The final prior used involves that analogous to placing a Beta prior on item uniqueness to prevent Heywood cases in item factor analysis (e.g., Bock et al., 1988; Mislevy, 1986). We show how to adapt this prior developed under multidimensional IRT to the current setting.[1]

Suppose we derived the LMPA item model using a probit rather than a logistic dis-

[1] A version of the LMPA model that omitted the Beta prior was presented at the 2013 International Meeting of the Psychometric Society. Omission of the Beta prior, a weaker prior on κ, and little troubleshooting of starting values tended to yield mediocre performance of the LMPA in simulations.

tribution. For example, underlying the observed response, y_{ij}, is a variable that is a function of a monotonic polynomial,

$$y_j^* = m(\theta, \boldsymbol{\lambda}_j) + \varepsilon_j = \sum_{q=1}^{2k+1} \lambda_{jq}\theta^q + \varepsilon_j \tag{8}$$

We may further assume that $\varepsilon_j \sim \mathcal{N}(0, \psi_j^2)$, and that $\theta \sim \mathcal{N}(0, \sigma^2)$, usually with $\sigma^2 = 1$. Dichotomous responses are produced from y_j^* via:

$$y_{ij} = \begin{cases} 1, & \text{if } y_{ij}^* \geq r_j \\ 0, & \text{if } y_{ij}^* < r_j \end{cases} \tag{9}$$

where r_j is the threshold for item j. The probability of a correct response under this model may be written as:

$$P^*(1|\theta) = c(\kappa_j) + \frac{1 - c(\kappa_j)}{\psi_j\sqrt{2\pi}} \int_{r_j}^{\infty} \exp\left\{ -(1/2)\left(\frac{r_j - m(\theta, \boldsymbol{\lambda}_j)}{\psi_j^2} \right)^2 \right\} dy_j^* \tag{10}$$

$$= c(\kappa_j) + (1 - c(\kappa_j))\Phi\left(-\frac{r_j - m(\theta, \boldsymbol{\lambda}_j)}{\psi_j^2} \right) \tag{11}$$

where Φ is the standard cumulative normal distribution function. This resembles standardized version of the normal ogive model with guessing (e.g., Bock et al., 1988), but with the linear predictor replaced by a monotonic polynomial. The goal in remedying Heywood cases in this situation involves preventing the unique variance ψ_j^2 from getting too small, or alternatively the steepness of the function implied by $m(\theta, \boldsymbol{\lambda}_j)$ from getting too large. If we assume a standard normal distribution for y_j^* and uncorrelated θ and ε_j, it follows that

$$\psi_j^2 = 1 - \text{var}(m(\theta, \boldsymbol{\lambda}_j)) \tag{12}$$

Thus, ψ_j^2 is not directly estimated, but is a function of $m(\theta, \boldsymbol{\lambda}_j)$. We may also approximate ψ_j^2 as a function of the parameters implied by the LMPA:

$$\psi^2 \approx \frac{1}{1 + (1/D^2)\text{var}(m(\theta, \mathbf{b}_j))} \tag{13}$$

where $D = 1.702$ is the usual scaling constant. Placing a Beta prior on ψ_j^2, such as $\pi(\psi_j^2) \sim B(p, q)^{-1}(\psi_j^2)^{p-1}(1 - \psi_j^2)^{q-1}$, where B is the Beta function and with $q = 1$, effectively prevents ψ_j^2 from being too small (e.g., Bock et al., 1988), and results in changes to the posterior mode for the item parameters in $m(\theta, \mathbf{b}_j) = m(\theta, \omega_j, \boldsymbol{\alpha}_j, \boldsymbol{\tau}_j)$. This strategy of developing weakly informative priors is not atypical in Bayesian inference, where one may choose to motivate the choice by imposing a prior on an alternative parameterization of the model (the item unique variance in this case), which then *induces* the desired prior on parameters of interest (the slopes). A typical choice for p is 1.5 (Cai et al., 2011), which we employ throughout this paper, and further details as to how the Beta prior changes the log-likelihood and derivatives are given in Appendix B.

3 Example: Large-scale state math assessment

To illustrate the LMPA item model, we utilized data from 10,000 Grade 4 students who provided responses to 44 dichotomously scored items on a state math assessment. The 3PL model and all LMPA models described in this section used priors as described in the previous section. As a preliminary check, use of a 3PL model for all items (AIC = 48,5581, BIC = 48,6533) indicated better fit than using a two-parameter logistic model for all items (AIC = 48,7290, BIC = 48,7924), and the mean pseudo-guessing parameter, $c(\kappa)$, for the 3PL model was approximately .17. We therefore concentrated on the 3PL as our baseline model.

3.1 Screening of item fit

The next step was to determine whether the LMPA model may be useful in improving model and/or item fit. For this task, we utilized $S - X^2$ using the version outlined by

Orlando and Thissen (2000). In brief, by conditioning on sum-score based groups, the test statistic detects whether observed counts of correct/incorrect are congruent with the expected counts based on the model. Specifically for item j,

$$S - X_j^2 = \sum_{t=1}^{v-1} N_t \frac{(O_{jt} - E_{jt})^2}{E_{jt}(1 - E_{jt})} \tag{14}$$

where the sum-score groups range from $0, 1, \ldots, v$ with v being the maximum observed sum-score, N_t is the number of respondents in sum-score group t, O_{jt} is the observed proportion of correct responses, and E_{jt} is the expected proportion of correct responses. E_{jt} in turn can be computed using the Lord-Wingersky algorithm (Lord & Wingersky, 1984) as described by Orlando and Thissen (2000). The statistic is compared to a central chi-square distribution with degrees of freedom of $df = t - 1 - z_j$ where z_j is the number of estimated parameters for item j. If adjacent sum score groups are collapsed due to low expected counts (usually less than 1), additional df adjustments are made.

The procedure for screening items for misfit for the empirical example, and in simulations, was as follows, using the baseline (3PL) model as the starting model.

1. Compute $S - X^2$ for all items in the current model.
2. Flag all items with misfit below a threshold (e.g., $p < .05$).
3. For all flagged items (if any), increase the order of the polynomial (e.g., if $k = 0$, change to $k = 1$; if $k = 1$ increase to $k = 2$).
4. If any items were changed as a result of Step 3, re-fit the model.

The Steps 1 through 4 may be repeated as many iterations as desired. For the current example and in simulations, we repeated this process only twice, meaning that at most the model was fit 3 times (3PL and two cycles of the above steps) and the maximum polynomial order for any fitted item was $k = 2$ (5th order). In addition, for Step 2, it is possible to employ either very liberal criteria in screening out items (e.g., no control

of Type I error or false discovery rate), or to employ a variety of corrective techniques. We experimented with using no correction (referred to hereafter as NC), and p-value adjustments using the Benjamini-Hochberg procedure (referred to hereafter as BH; Benjamini & Hochberg, 1995), which is sometimes advocated for use in IRT contexts as a high-power alternative to the Bonferroni method (Thissen, Steinberg, & Kuang, 2002). While we may expect the NC approach to lead to overfitting, it will have higher power and we later examine in simulations whether such overfitting has averse consequences.

3.2 LMPA Results

$S - X^2$ initially flagged 11 items as poorly fitting under NC and 7 items when using the BH correction. Thus, roughly 16% to 25% of items may have poor fit. Flagging and fitting items twice under NC resulted in 6 items modeled with $k = 1$ and 6 items with $k = 2$ using the LMPA model. Under BH, these numbers were predictably lower, with 5 items as $k = 1$ and 2 items as $k = 2$. Both approaches led to fewer items having poor fit according to $S - X^2$. For instance, the final NC model had 6 items with poor fit and the final BH model had only 2. As shown in Table 2, both final NC and BH models also improved AIC, with the NC model indicating slightly better fit. However, BIC actually preferred the 3PL model.

We interpret these results as meaning that there is some promise for the LMPA in reducing the number of poorly fitting items according to $S - X^2$. The information criterion results are mixed in terms of whether the overall final model is an improvement, with AIC suggesting an improvement, but a higher penalization for more model parameters under BIC suggesting overfitting. It is therefore difficult to guess whether the LMPA is substantially improving IRF recovery. We next turn to simulation results to further test the potential impact of the LMPA and our modeling procedure.

4 Simulation

4.1 Method

We performed a small simulation study to test the performance of the LMPA IRT model and the procedure we employed in using $S - X^2$ to flag candidate items. We used a 2 (% non-standard IRF: 20% vs. 40% of items) $\times$ 2 (N: 1,000 vs. 5,000) overall design. In all cases, we used 40 items, generated latent proficiencies, θ, from a standard normal distribution, and performed 100 replications per cell of the design. Studied conditions were thus chosen to partially overlap with conditions found with the empirical example.

Most items were 3PL items (LMPA with $k = 0$) with item parameters (κ, ω, and δ) randomly drawn in each replication from a multivariate normal distribution that matched the mean and covariance of the all 3PL model in the previous section. To avoid too many extreme items, parameter draws with any item greater than 1.65 SD away from its mean were discarded and a new set was drawn. The remaining items (8 or 16 items, depending on condition), were randomly constructed from a mixture of normal cumulative distribution functions (CDF) and lower asymptotes. Specifically, $p_1(g_1 + (1-g_1)\Phi(\theta|\mu_1, \sigma_1^2)) + p_2(g_2 + (1-g_2)\Phi(\theta|\mu_2, \sigma_2^2)) + p_3(g_3 + (1-g_3)\Phi(\theta|\mu_3\sigma_3^2))$, with p_1 and $p_2 \sim$ unif$(.2, .4)$, $p_3 = 1 - p_1 - p_2$, $\mu_1 \sim \mathcal{N}(-1.5, .4^2)$, $\mu_2 \sim \mathcal{N}(1.5, .4^2)$, $\mu_3 \sim \mathcal{N}(0, .4^2)$, all g parameters drawn from unif$(.1, .3)$, and σ_1, σ_2, and σ_3 independently drawn from a log-normal distribution, $\ln \mathcal{N}(-1.03, .22)$. Here we use Φ to denote the cumulative distribution function along θ for a given mean and variance.

To each generated dataset, we mimicked the steps taken in our empirical example. That is, we initially fit a 3PL model, followed by flagging items using NC or BH, and fitting flagged items with higher-order polynomials. We repeated re-fitting twice for both NC and BH, resulting in final NC and BH models. Thus, we are interested in comparing the 3PL to the final NC and BH models. Numerical integration necessary for estimation was done with 49 equally spaced quadrature points between -5 and 5 across θ. The maximum number of M-step iterations was set at 50, and the maximum number

of EM cycles was set at 500. Model estimation terminated if item parameters from one EM cycle to the next differed by 1e-4 or less.

To further mimic a real data analysis situation, we never started estimation at the true model parameters, but rather at $\kappa = -1$, $\delta = 0$, $\omega = 0$, $\alpha = 0$, and $\tau = -1$ for all items. However, to automate troubleshooting of estimation problems, we employed the following ad-hoc procedure for all replications. Initial starting values for item parameters were further tweaked by using the final estimates of a model run with very few maximum M-step iterations and EM cycles (2 and 5, respectively). We then commenced with estimation as outlined in the previous paragraph. If the maximum number of EM iterations was reached, or if ridging the Hessian ever failed to yield an invertible matrix, new starting values were attempted by normalizing sum scores across simulees for use in the complete-data likelihood for each item. The item parameter estimates based on this complete-data analysis were then used as starting values.

4.2 Results

Overall it was our hope and expectation that the procedure of using $S - X^2$ would tend to correctly flag non-standard IRFs (using either NC or BH), which could in turn be modelled by the LMPA with higher-order polynomials. This may result in better recovery of individual IRFs.

4.2.1 Ability of $S - X^2$ to detect non-standard IRFs

In general, power to detect non-standard IRFs was slightly less than anticipated. Table 3 displays power and false positive rates based on the final NC and BH models for all cells in the design. Power is the proportion of items with non-standard IRFs that were fitted with the LMPA model with $k > 0$, whereas the false positive rate is the proportion of items with a true 3PL shape fitted by the LMPA model with $k > 0$. Both were computed across all replications in each cell. Thus, power to detect non-standard IRFs only reaches a maximum of .36 or .33 for NC and .14 for BH (both when $N = 5,000$). This amounts to correctly detecting approximately 3/8 items for NC and 1/8 items for

BH when 20% of items have non-standard IRFs, or 5 to 6/16 items for NC and 2/16 items for BH when 40% of items have non-standard IRFs.

While we do not immediately have a full explanation, we note that (Orlando & Thissen, 2003) found the lowest power for $S - X^2$ for items that had a non-standard bend or a plateau in the middle of the IRF - visually similar to the items we generated - versus items that had non-monotonicity or an omitted asymptote. Since non-standard IRFs were generated via a random process, it is not necessarily the case that all items would in fact look very extreme. This low power rate may not necessarily be problematic, however, to the extent that both approaches may flag the worst-fitting items and flag non-standard items at a higher rate than 3PL items.

4.2.2 Recovery of IRFs

We assessed recovery of IRFs by using Root Integrated Mean Square Error (RIMSE; e.g., Liang, 2007; Ramsay, 1991), defined as the following for item j:

$$\text{RIMSE}_j = \left(\frac{\sum_{q=1}^{Q} (\hat{P}_j(1|\theta_q) - P_j(1|\theta_q))^2 \phi(\theta_q)}{\sum_{q=1}^{Q} \phi(\theta_q)} \right)^{1/2} \times 100 \tag{15}$$

where $\hat{P}(1|\theta)$ and $P(1|\theta)$ are short-hand for the estimated and true response function for the correct response, ϕ is a standard normal density function, and the sum is across a series of Q equally spaced quadrature points along θ (-5 to 5 in .1 increments). RIMSE thus represents the root of a weighted average squared discrepancy between the true and estimated response functions, with more weight given to discrepancies towards the middle of the θ distribution. Lower values indicate better IRF recovery. This index was computed for all estimated items and was aggregated across replications and items where necessary.

Since NC and BH approaches did not always suggest poorly fitting items for each replication, we separately compared each of NC and BH to the 3PL model, focusing only on those replications where the final BH and NC models differed from the 3PL. Overall,

the NC model slightly improved RIMSE versus the 3PL model across all conditions (see Table 4), with gains more prominent at the larger sample size ($N = 5000$), perhaps in part to higher power to flag non-standard IRFs. Examining items where the true underlying IRF was either a 3PL or a non-standard item revealed that all of these gains were because of improvement in recovering non-standard items. In fact, recovery of 3PL items was slightly worse for the NC model, though perhaps negligibly so.

Overall, the BH model also slightly improved RIMSE versus the 3PL model across all studied conditions (see Table 5), in a similar pattern to NC. Gains in RIMSE were also due to better modeling of non-standard items. Albeit based on a much smaller number of replications, use of BH appeared to have no averse impact on the recovery of 3PL item IRFs, in slight contrast to using NC.

5 Discussion

We have presented a new IRT model, the logistic function of a monotonic polynomial with asymptote (LMPA), that can account for guessing as does the 3PL, but has a more flexible IRF shape. We proposed a possible data generating mechanism that may yield such non-standard IRFs - heterogeneous populations whose IRFs are mixed together in generating the item responses. Finally, we tested use of $S - X^2$ to flag candidate items for use with the LMPA model in an empirical example and through simulations.

Our empirical example suggests that use of $S - X^2$ and LMPA can improve the number of well-fitting items without it being necessary to employ a computationally intensive AIC or BIC step-wise approach. For instance, the number of poorly fitting items was reduced by approximately half or better, depending on whether NC or BH p-values under $S - X^2$ were examined. Such initially poor-fitting items may be those that show instructional sensitivity in cases where curriculum and instruction is not implemented in a fully standard way across the entire population. Our procedure thus serves the dual purpose of identifying potential items, and allowing retention of such items (which can be expensive to develop) in a large-scale assessment if practice dictates that poor fitting

items are flagged for possible deactivation.

Our simulation results suggest that the procedure utilized on our empirical example is sound and can lead to better IRF recovery. This was especially the case with a higher sample size, and to a certain extent cases where more items had true non-standard IRFs. Interestingly, however, it is difficult to make a strong recommendation on whether no p-value correction for $S - X^2$ should be employed or whether to use the Benjamini-Hochberg (Benjamini & Hochberg, 1995) correction. Both yielded comparable gains in RIMSE. If parsimony is preferred, the BH approach would yield fewer items modelled using the LMPA approach.

Appendix A. Derivatives for the LMPA model.

We used complete first- and second-order derivatives of the LMPA model in the M-step of the EM algorithm. We drop item and respondent subscripts and adopt the following short-hand notation for the response function in Equation 3:

$$P = c(\kappa) + \frac{1 - c(\kappa)}{1 + \exp(-(\delta + m(\theta)))} \tag{16}$$

where $m(\theta)$ is short-hand for the monotonic polynomial in Equation 4. We also define $Q = 1 - P$ as the probability of an incorrect response. The complete-data log-likelihood for a single item and response can then be written as:

$$l = y \log(P) + (1 - y) \log(1 - P) \tag{17}$$

where y is the observed response. Differentiating (17) with respect to any item parameter, η, leads to:

$$\frac{\partial l}{\partial \eta} = \frac{y - P}{PQ} \frac{\partial P}{\partial \eta} \tag{18}$$

First-order derivatives are obtained by substitution for $\frac{\partial P}{\partial \eta}$:

$$\begin{aligned}
\frac{\partial P}{\partial \kappa} &= c(\kappa)(1 - c(\kappa))(1 - c(m(\theta))) \\
\frac{\partial P}{\partial \delta} &= (1 - c(\kappa))c(m(\theta))(1 - c(m(\theta))) \\
\frac{\partial P}{\partial \omega} &= (1 - c(\kappa))c(m(\theta))(1 - c(m(\theta))) \frac{\partial m(\theta)}{\partial \mathbf{a}} \frac{\partial \mathbf{a}}{\partial \omega} \\
\frac{\partial P}{\partial \alpha_s} &= (1 - c(\kappa))c(m(\theta))(1 - c(m(\theta))) \frac{\partial m(\theta)}{\partial \mathbf{a}} \frac{\partial \mathbf{a}}{\partial \alpha_s} \\
\frac{\partial P}{\partial \tau_s} &= (1 - c(\kappa))c(m(\theta))(1 - c(m(\theta))) \frac{\partial m(\theta)}{\partial \mathbf{a}} \frac{\partial \mathbf{a}}{\partial \tau_s}
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial m(\theta)}{\partial \mathbf{a}} &= \begin{bmatrix} \theta & \frac{1}{2}\theta^2 & \frac{1}{3}\theta^3 & \cdots & \frac{1}{2k+1}\theta^{2k+1} \end{bmatrix} \\
\frac{\partial \mathbf{a}}{\partial \omega} &= \mathbf{T}_k \mathbf{T}_{k-1} \cdots \mathbf{T}_2 \mathbf{T}_1 \exp(\omega) \\
\frac{\partial \mathbf{a}}{\partial \alpha_s} &= \mathbf{T}_k \mathbf{T}_{k-1} \cdots \frac{\partial \mathbf{T}_s}{\partial \alpha_s} \cdots \mathbf{T}_2 \mathbf{T}_1 \exp(\omega) \\
\frac{\partial \mathbf{a}}{\partial \tau_s} &= \mathbf{T}_k \mathbf{T}_{k-1} \cdots \frac{\partial \mathbf{T}_s}{\partial \tau_s} \cdots \mathbf{T}_2 \mathbf{T}_1 \exp(\omega)
\end{aligned}$$

and derivatives of the matrices $\mathbf{T}$ are as follows:

$$\frac{\partial \mathbf{T}_s}{\partial \alpha_{js}} = \begin{bmatrix}
0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
-2 & 0 & 0 & \cdots & 0 & 0 & 0 \\
2\alpha_{js} & -2 & 0 & \cdots & 0 & 0 & 0 \\
0 & 2\alpha_{js} & -2 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 2\alpha_{js} & -2 & 0 \\
0 & 0 & 0 & \cdots & 0 & 2\alpha_{js} & -2 \\
0 & 0 & 0 & \cdots & 0 & 0 & 2\alpha_{js}
\end{bmatrix}$$

$$\frac{\partial \mathbf{T}_s}{\partial \tau_{js}} = \begin{bmatrix}
0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
\exp(\tau_{js}) & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & \exp(\tau_{js}) & 0 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & \exp(\tau_{js}) & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & \exp(\tau_{js}) & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & \exp(\tau_{js})
\end{bmatrix}$$

To obtain second-order derivatives, we add subscripts to η and differentiate again,

$$\begin{aligned}\frac{\partial^2 l}{\partial\eta_1\partial\eta_2} &= \frac{\partial}{\partial\eta_1}\left[\frac{\partial P}{\partial\eta_2}\frac{y-P}{PQ}\right] \\ &= \frac{\partial^2 P}{\partial\eta_1\partial\eta_2}\left[\frac{y-P}{PQ}\right] + \frac{\partial P}{\partial\eta_2}\frac{\partial P}{\partial\eta_1}\left[-\frac{1}{PQ} - \frac{(y-P)(Q-P)}{(PQ)^2}\right]\end{aligned}$$

Again, we substitute for each type of item parameter:

$$\begin{aligned}
\frac{\partial^2 P}{\partial\kappa\partial\kappa} &= c(\kappa)(1-c(\kappa))(1-2c(\kappa))(1-c(m(\theta))) \\
\frac{\partial^2 P}{\partial\kappa\partial\delta} &= -c(\kappa)(1-c(\kappa))c(m(\theta))(1-c(m(\theta))) \\
\frac{\partial^2 P}{\partial\kappa\partial\omega} &= -c(\kappa)(1-c(\kappa))c(m(\theta))(1-c(m(\theta)))\frac{\partial m(\theta)}{\partial\mathbf{a}}\frac{\partial\mathbf{a}}{\partial\omega} \\
\frac{\partial^2 P}{\partial\kappa\partial\alpha_s} &= -c(\kappa)(1-c(\kappa))c(m(\theta))(1-c(m(\theta)))\frac{\partial m(\theta)}{\partial\mathbf{a}}\frac{\partial\mathbf{a}}{\partial\alpha_s} \\
\frac{\partial^2 P}{\partial\kappa\partial\tau_s} &= -c(\kappa)(1-c(\kappa))c(m(\theta))(1-c(m(\theta)))\frac{\partial m(\theta)}{\partial\mathbf{a}}\frac{\partial\mathbf{a}}{\partial\tau_s} \\
\frac{\partial^2 P}{\partial\delta\partial\delta} &= (1-c(\kappa))c(m(\theta))(1-c(m(\theta)))(1-2(c(m(\theta)))) \\
\frac{\partial^2 P}{\partial\delta\partial\omega} &= (1-c(\kappa))c(m(\theta))(1-c(m(\theta)))(1-2(c(m(\theta))))\frac{\partial m(\theta)}{\partial\mathbf{a}}\frac{\partial\mathbf{a}}{\partial\omega} \\
\frac{\partial^2 P}{\partial\delta\partial\alpha_s} &= (1-c(\kappa))c(m(\theta))(1-c(m(\theta)))(1-2(c(m(\theta))))\frac{\partial m(\theta)}{\partial\mathbf{a}}\frac{\partial\mathbf{a}}{\partial\alpha_s} \\
\frac{\partial^2 P}{\partial\delta\partial\tau_s} &= (1-c(\kappa))c(m(\theta))(1-c(m(\theta)))(1-2(c(m(\theta))))\frac{\partial m(\theta)}{\partial\mathbf{a}}\frac{\partial\mathbf{a}}{\partial\tau_s} \\
\frac{\partial^2 P}{\partial\omega\partial\omega} &= (1-c(\kappa))c(m(\theta))(1-c(m(\theta)))\left[(1-2c(m(\theta)))\frac{\partial m(\theta)}{\partial\omega}\frac{\partial m(\theta)}{\partial\omega} + \frac{\partial^2 m(\theta)}{\partial\omega\partial\omega}\right] \\
\frac{\partial^2 P}{\partial\omega\partial\alpha_s} &= (1-c(\kappa))c(m(\theta))(1-c(m(\theta)))\left[(1-2c(m(\theta)))\frac{\partial m(\theta)}{\partial\alpha_s}\frac{\partial m(\theta)}{\partial\omega} + \frac{\partial^2 m(\theta)}{\partial\omega\partial\alpha_s}\right] \\
\frac{\partial^2 P}{\partial\omega\partial\tau_s} &= (1-c(\kappa))c(m(\theta))(1-c(m(\theta)))\left[(1-2c(m(\theta)))\frac{\partial m(\theta)}{\partial\tau_s}\frac{\partial m(\theta)}{\partial\omega} + \frac{\partial^2 m(\theta)}{\partial\omega\partial\tau_s}\right] \\
\frac{\partial^2 P}{\partial\alpha_s\partial\alpha_s} &= (1-c(\kappa))c(m(\theta))(1-c(m(\theta)))\left[(1-2c(m(\theta)))\frac{\partial m(\theta)}{\partial\alpha_s}\frac{\partial m(\theta)}{\partial\alpha_s} + \frac{\partial^2 m(\theta)}{\partial\alpha_s\partial\alpha_s}\right] \\
\frac{\partial^2 P}{\partial\alpha_s\partial\alpha_t} &= (1-c(\kappa))c(m(\theta))(1-c(m(\theta)))\left[(1-2c(m(\theta)))\frac{\partial m(\theta)}{\partial\alpha_s}\frac{\partial m(\theta)}{\partial\alpha_t} + \frac{\partial^2 m(\theta)}{\partial\alpha_s\partial\alpha_t}\right] \\
\frac{\partial^2 P}{\partial\alpha_s\partial\tau_s} &= (1-c(\kappa))c(m(\theta))(1-c(m(\theta)))\left[(1-2c(m(\theta)))\frac{\partial m(\theta)}{\partial\alpha_s}\frac{\partial m(\theta)}{\partial\tau_s} + \frac{\partial^2 m(\theta)}{\partial\alpha_s\partial\tau_s}\right]
\end{aligned}$$

$$\frac{\partial^2 P}{\partial \alpha_s \partial \tau_t} = (1-c(\kappa))c(m(\theta))(1-c(m(\theta)))\left[(1-2c(m(\theta)))\frac{\partial m(\theta)}{\partial \alpha_s}\frac{\partial m(\theta)}{\partial \tau_t} + \frac{\partial^2 m(\theta)}{\partial \alpha_s \partial \tau_t}\right]$$
$$\frac{\partial^2 P}{\partial \tau_s \partial \tau_s} = (1-c(\kappa))c(m(\theta))(1-c(m(\theta)))\left[(1-2c(m(\theta)))\frac{\partial m(\theta)}{\partial \tau_s}\frac{\partial m(\theta)}{\partial \tau_s} + \frac{\partial^2 m(\theta)}{\partial \tau_s \partial \tau_s}\right]$$
$$\frac{\partial^2 P}{\partial \tau_s \partial \tau_t} = (1-c(\kappa))c(m(\theta))(1-c(m(\theta)))\left[(1-2c(m(\theta)))\frac{\partial m(\theta)}{\partial \tau_t}\frac{\partial m(\theta)}{\partial \tau_s} + \frac{\partial^2 m(\theta)}{\partial \tau_s \partial \tau_t}\right]$$

Where second-order derivatives for $m(\theta)$ are as follows:

$$\frac{\partial^2 m(\theta)}{\partial \omega \partial \omega} = \frac{\partial m(\theta)}{\partial \mathbf{a}}\mathbf{T}_k\mathbf{T}_{k-1}\cdots\mathbf{T}_2\mathbf{T}_1\exp(\omega)$$
$$\frac{\partial^2 m(\theta)}{\partial \omega \partial \alpha_s} = \frac{\partial m(\theta)}{\partial \mathbf{a}}\mathbf{T}_k\mathbf{T}_{k-1}\cdots\frac{\partial \mathbf{T}_s}{\partial \alpha_s}\cdots\mathbf{T}_2\mathbf{T}_1\exp(\omega)$$
$$\frac{\partial^2 m(\theta)}{\partial \omega \partial \tau_s} = \frac{\partial m(\theta)}{\partial \mathbf{a}}\mathbf{T}_k\mathbf{T}_{k-1}\cdots\frac{\partial \mathbf{T}_s}{\partial \tau_s}\cdots\mathbf{T}_2\mathbf{T}_1\exp(\omega)$$
$$\frac{\partial^2 m(\theta)}{\partial \alpha_s \partial \alpha_s} = \frac{\partial m(\theta)}{\partial \mathbf{a}}\mathbf{T}_k\mathbf{T}_{k-1}\cdots\frac{\partial^2 \mathbf{T}_s}{\partial \alpha_s \partial \alpha_s}\cdots\mathbf{T}_2\mathbf{T}_1\exp(\omega)$$
$$\frac{\partial^2 m(\theta)}{\partial \alpha_s \partial \alpha_t} = \frac{\partial m(\theta)}{\partial \mathbf{a}}\mathbf{T}_k\mathbf{T}_{k-1}\cdots\frac{\partial \mathbf{T}_s}{\partial \alpha_s}\cdots\frac{\partial \mathbf{T}_t}{\partial \alpha_t}\cdots\mathbf{T}_2\mathbf{T}_1\exp(\omega)$$
$$\frac{\partial^2 m(\theta)}{\partial \alpha_s \partial \tau_s} = \frac{\partial m(\theta)}{\partial \mathbf{a}}\mathbf{T}_k\mathbf{T}_{k-1}\cdots\frac{\partial^2 \mathbf{T}_s}{\partial \alpha_s \partial \tau_s}\cdots\mathbf{T}_2\mathbf{T}_1\exp(\omega)$$
$$\frac{\partial^2 m(\theta)}{\partial \alpha_s \partial \tau_t} = \frac{\partial m(\theta)}{\partial \mathbf{a}}\mathbf{T}_k\mathbf{T}_{k-1}\cdots\frac{\partial \mathbf{T}_s}{\partial \alpha_s}\cdots\frac{\partial \mathbf{T}_t}{\partial \tau_t}\cdots\mathbf{T}_2\mathbf{T}_1\exp(\omega)$$
$$\frac{\partial^2 m(\theta)}{\partial \tau_s \partial \tau_s} = \frac{\partial m(\theta)}{\partial \mathbf{a}}\mathbf{T}_k\mathbf{T}_{k-1}\cdots\frac{\partial^2 \mathbf{T}_s}{\partial \tau_s \partial \tau_s}\cdots\mathbf{T}_2\mathbf{T}_1\exp(\omega)$$
$$\frac{\partial^2 m(\theta)}{\partial \tau_s \partial \tau_t} = \frac{\partial m(\theta)}{\partial \mathbf{a}}\mathbf{T}_k\mathbf{T}_{k-1}\cdots\frac{\partial \mathbf{T}_s}{\partial \tau_s}\cdots\frac{\partial \mathbf{T}_t}{\partial \tau_t}\cdots\mathbf{T}_2\mathbf{T}_1\exp(\omega)$$

Finally, second-order derivatives of the matrices $\mathbf{T}$ are the following:

$$\frac{\partial^2 \mathbf{T}_s}{\partial \alpha_s \partial \alpha_s} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 2 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 2 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 2 \end{bmatrix}$$

$$\frac{\partial^2 \mathbf{T}_s}{\partial \alpha_s \partial \tau_s} = \mathbf{0}$$

$$\frac{\partial^2 \mathbf{T}_s}{\partial \tau_s \partial \tau_s} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \exp(\tau_s) & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \exp(\tau_s) & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \exp(\tau_s) & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \exp(\tau_s) & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \exp(\tau_s) \end{bmatrix}$$

A Appendix B. Beta prior with the LMPA model.

As in the previous Appendix, we adopt the short-hand notation of $m(\theta)$ for the monotonic polynomial. According to Equation 13, the unique item variance is approximately

the following under the LMPA model:

$$\psi^2 \approx \frac{1}{1 + (1/D^2)\text{var}(m(\theta))} \tag{19}$$

Since we are considering a single latent trait, the variance of the monotonic polynomial is the following:

$$\text{var}(m(\theta)) = \mathbf{b}'\mathbf{\Gamma}\mathbf{b} \tag{20}$$

where $\mathbf{b}$ are the coefficients for the polynomial, and $\mathbf{\Gamma}$ is a symmetric matrix:

$$\mathbf{\Gamma} = \begin{bmatrix} \text{var}(\theta) & & & \\ \text{cov}(\theta, \theta^2) & \text{var}(\theta^2) & & \\ \vdots & \vdots & \ddots & \\ \text{cov}(\theta, \theta^{2k+1}) & \text{cov}(\theta^2, \theta^{2k+1}) & \dots & \text{var}(\theta^{2k+1}) \end{bmatrix} \tag{21}$$

The elements of $\mathbf{\Gamma}$ can be obtained simply by computing the moments of the distribution for θ, which in our application is a standard normal distribution. Addition of a Beta prior ridges the complete-data log-likelihood by the following:

$$(-1)\log(B(p, q)) + (p-1)\log(\psi^2) \tag{22}$$

Differentiating this quantity with respect to a typical parameter, η, we have:

$$\frac{\partial (p-1)\log(\psi^2)}{\partial \eta} = (p-1)\frac{1}{\psi^2}\frac{\partial \psi^2}{\partial \eta} \tag{23}$$

$$= \frac{-(2/D^2)(p-1)}{1 + (1/D^2)\mathbf{b}'\mathbf{\Gamma}\mathbf{b}}\frac{\partial \mathbf{a}}{\partial \eta}\frac{\partial \mathbf{b}}{\partial \mathbf{a}}\mathbf{\Gamma}\mathbf{b} \tag{24}$$

where $\frac{\partial \mathbf{a}}{\partial \eta}$ for each type of model parameter are given in Appendix A, and $\frac{\partial \mathbf{b}}{\partial \mathbf{a}}$ is a diagonal matrix:

$$\frac{\partial \mathbf{b}}{\partial \mathbf{a}} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1/2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1/(2k+1) \end{bmatrix}$$

Thus, when a Beta prior is present, complete-data first-order derivatives are adjusted by the quantity in (23). Second-order derivatives are adjusted by the quantity on the following page, obtained by differentiating (23) a second time.

$$\frac{\partial^2 (p-1)\log(\psi^2)}{\partial \eta_1 \partial \eta_2} \tag{25}$$

$$= -(2/D^2)(p-1)\psi^2 \left\{ -(2/D^2)\psi^2 \left(\frac{\partial \mathbf{a}}{\partial \eta_1} \frac{\partial \mathbf{b}}{\partial \mathbf{a}} \mathbf{\Gamma b} \right) \left(\frac{\partial \mathbf{a}}{\partial \eta_2} \frac{\partial \mathbf{b}}{\partial \mathbf{a}} \mathbf{\Gamma b} \right) + \frac{\partial^2 \mathbf{a}}{\partial \eta_1 \partial \eta_2} \frac{\partial \mathbf{b}}{\partial \mathbf{a}} \mathbf{\Gamma b} + \frac{\partial \mathbf{a}}{\partial \eta_1} \left(\frac{\partial \mathbf{b}}{\partial \mathbf{a}} \mathbf{\Gamma} \right)' \left(\frac{\partial \mathbf{a}}{\partial \eta_2} \right)' \right\} \tag{26}$$

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289-300.

Birnbaum, A. (1968). Some latent trait models. In F. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (p. 395-479). Reading, MA: Addison-Wesley.

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*, 261–280.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized item bifactor analysis. *Psychological Methods*, *16*(3), 221-248.

Duncan, K. A., & MacEachern, S. N. (2008). Nonparametric Bayesian modelling for item response. *Statistical Modelling*, *8*(1), 41-66.

Duncan, K. A., & MacEachern, S. N. (2013). Nonparametric Bayesian modeling of item response curves with a three-parameter logistic prior mean. In M. C. Edwards & R. C. MacCallum (Eds.), *Current topics in the theory and application of latent variable models* (pp. 108–125). New York, NY: Routledge.

Elphinstone, C. D. (1985). *A method of distribution and density estimation*. Unpublished doctoral dissertation, University of South Africa.

Falk, C. F., & Cai, L. (in press). Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis. *Psychometrika*.

Liang, L. (2007). *A semi-parametric approach to estimating item response functions*. Unpublished doctoral dissertation, Department of Psychology, The Ohio State University.

Liang, L., & Browne, M. W. (2015). A quasi-parametric method for fitting flexible item response functions. *Journal of Educational and Behavioral Statistics*, *40*, 5-34.

Lord, F., & Wingersky, M. (1984). Comparison of IRT true-score and equipercentile observed score "equatings.". *Applied Psychological Measurement*, *8*, 453-461.

Mislevy, R. J. (1986). Bayes model estimation in item response models. *Psychometrika*, *51*(2), 177-195.

Miyazaki, K., & Hoshino, T. (2009). A bayesian semiparametric item response model with drichlet process priors. *Psychometrika*, *74*(3), 375-393.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of $S - X^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*, 289–298.

Qin, L. (1998). *Nonparametric Bayesian models for item response data*. Unpublished doctoral dissertation, The Ohio State University.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*(4), 611-630.

Ramsay, J. O., & Abrahamowicz, M. (1989). Binomial regression with monotone splines: A psychometric application. *Journal of the American Statistical Association*, *84*(408), 906-915.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271-282.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, *27*, 77–83.

Table 1: Item Parameters from Example IRFs

Parameter	Item 2 ($k = 0$ / 3PL)	Item 24 ($k = 0$ / 3PL)	Item 2 ($k = 2$)	Item 24 ($k = 1$)
κ	-1.91	-2.14	-1.81	-1.59
δ	-2.19	.77	-2.25	.80
ω	.70	-.54	1.14	-.84
α_1			.59	.40
α_2			.35	
τ_1			.25	-1.20
τ_2			-6.47	

Note. IRF = item response function; 3PL = three parameter logistic.

Table 2: Model overview of empirical example

	3PL	NC Final Model	BH Final Model
# Parameters	132	168	150
$-2 \log L$	485317	485111	485199
AIC	485581	485447	485499
BIC	486533	486659	486580

Note. 3PL = three parameter logistic; NC = no correction to $S - X^2$ p-values; BH = Benjamini-Hochberg correction to $S - X^2$ p-values.

Table 3: $S - X^2$ detection rates for the simulation study

	20% non-standard IRFs		40% non-standard IRFs	
	$N = 1,000$	$N = 5,000$	$N = 1,000$	$N = 5,000$
Power				
NC	.101	.361	.111	.336
BH	.011	.144	.009	.144
False positive				
NC	.052	.055	.059	.082
BH	.003	.004	.002	.011

Note. IRF = item response function; NC = no correction to $S - X^2$ p-values; BH = Benjamini-Hochberg correction to $S - X^2$ p-values.

Table 4: IRF recovery (RIMSE) for NC from the simulation study

	# rep.	Overall RIMSE	3PL items	Non-standard items
$N = 1,000$				
20% non-standard IRFs				
3PL Model	89	2.67	2.17	4.64
NC Model	89	2.66	2.24	4.48
40% non-standard IRFs				
3PL Model	95	3.16	2.14	4.67
NC Model	95	3.12	2.21	4.48
$N = 5,000$				
20% non-standard IRFs				
3PL Model	97	1.62	1.03	4.01
NC Model	97	1.50	1.06	3.23
40% non-standard IRFs				
3PL Model	100	2.22	1.04	3.99
NC Model	100	1.96	1.09	3.26

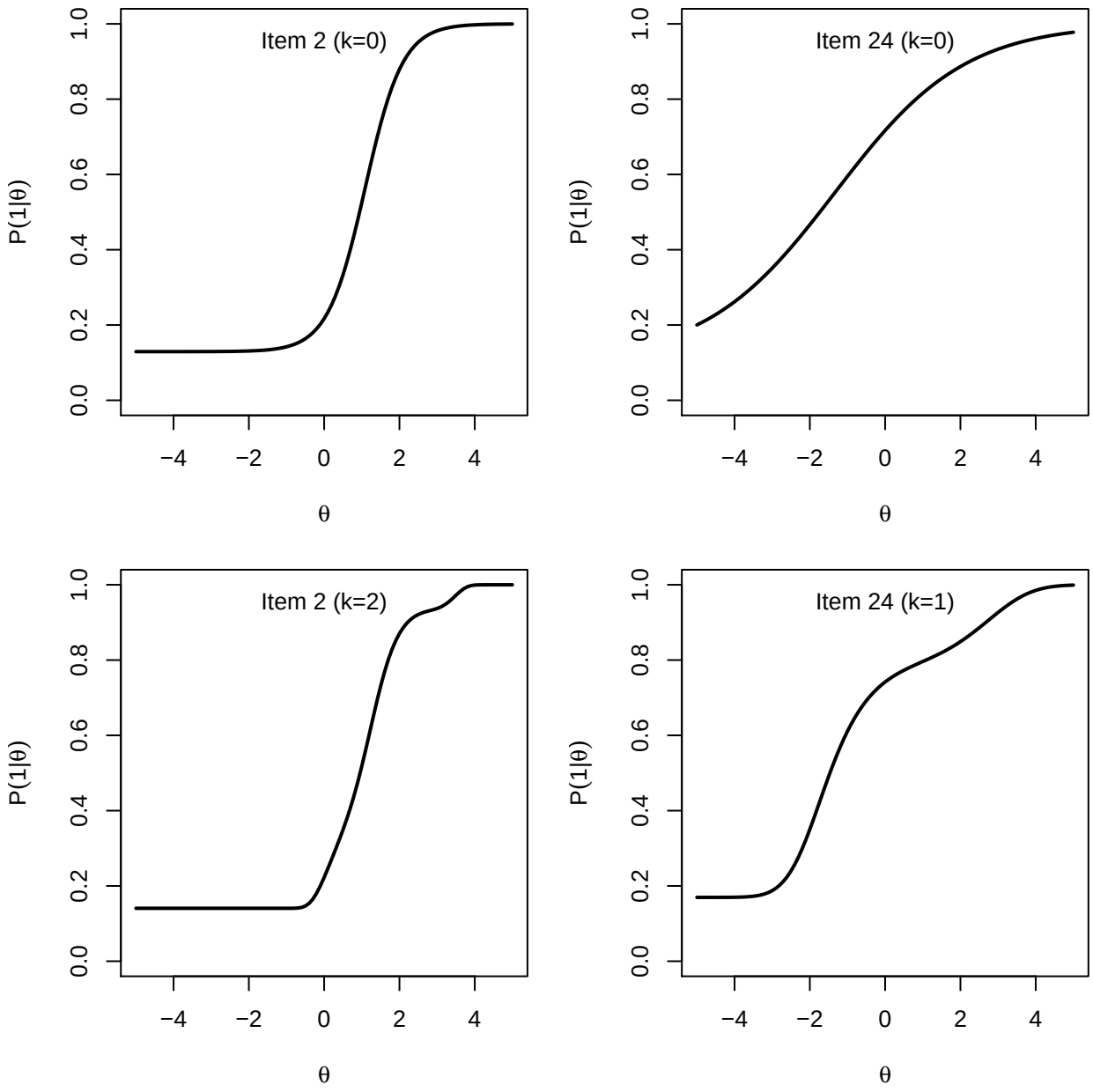
Note. IRF = item response function; RIMSE = Root integrated mean square error; 3PL = three parameter logistic; NC = no correction to $S - X^2$ p-values; BH = Benjamini-Hochberg correction to $S - X^2$ p-values.

Table 5: IRF recovery (RIMSE) for BH from the simulation study

	# rep.	Overall RIMSE	3PL items	Non-standard items
$N = 1,000$				
20% non-standard IRFs				
3PL Model	15	2.71	2.18	4.84
BH Model	15	2.66	2.19	4.54
40% non-standard IRFs				
3PL Model	16	3.18	2.16	4.71
BH Model	16	3.13	2.18	4.56
$N = 5,000$				
20% non-standard IRFs				
3PL Model	67	1.65	1.02	4.14
BH Model	67	1.53	1.02	3.58
40% non-standard IRFs				
3PL Model	89	2.23	1.04	4.01
BH Model	89	2.04	1.03	3.55

Note. IRF = item response function; RIMSE = Root integrated mean square error; 3PL = three parameter logistic; NC = no correction to $S - X^2$ p-values; BH = Benjamini-Hochberg correction to $S - X^2$ p-values.

Figure 1: Example IRFs



Note. The top two panels are response functions fit to math assessment data using the 3PL item model, whereas the bottom panels are the same items fit with the LMPA model.

Figure 2: Mixture of IRFs

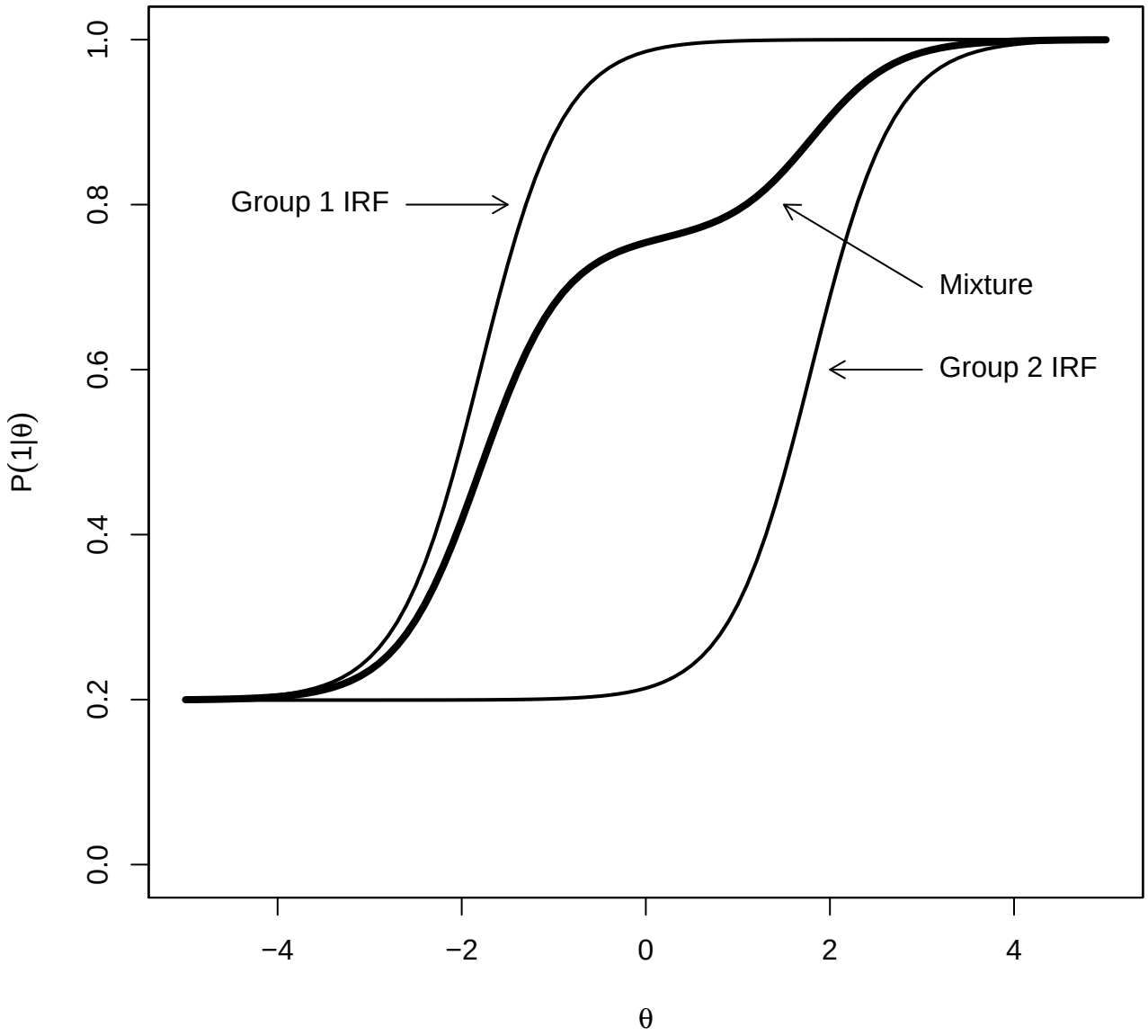
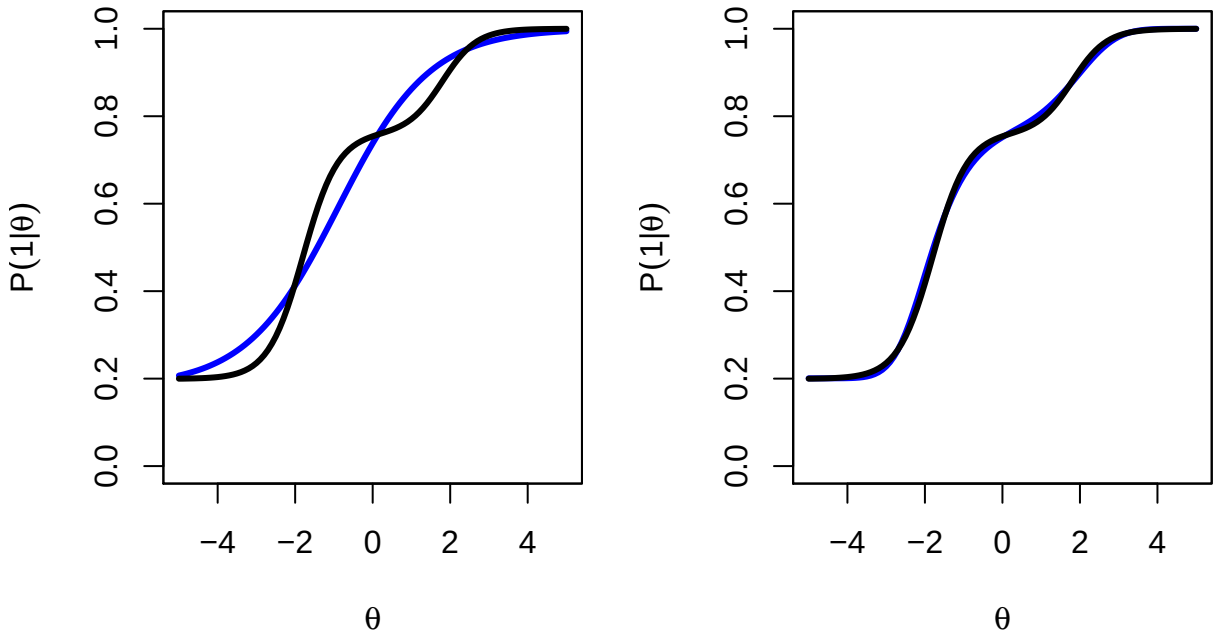


Figure 3: Best fitting 3PL and LMPA functions to mixture IRF



Note. Black line is mixture IRF. Left panel includes best fitting 3PL (blue), and right panel includes best fitting LMPA line (blue) using $k = 1$.