Title:
Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis

Authors:
Carl F. Falk
Li Cai

Journal publication date:
2016

Published in:
Psychometrika, 81(2), 434-460

IES grant information:
Grant number R305D140046
Funded by National Center for Education Research (NCER)

Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis

Carl F. Falk
University of California, Los Angeles

Li Cai
CRESST/University of California, Los Angeles

Manuscript in press at *Psychometrika*

This research is supported by a Social Sciences and Humanities Research Council of Canada Post-Doctoral Fellowship awarded to Carl F. Falk. Li Cai's research is partially supported by grants from the Institute of Education Sciences (R305B080016 and R305D100039) and grants from the National Institute on Drug Abuse (R01DA026943 and R01DA030466).

Address all correspondence to: Carl F. Falk, Graduate School of Education & Information Studies, UCLA, Los Angeles, CA, USA 90095-1521. Email: cffalk@gmail.com. Phone: 562.221.7538.

Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis

Abstract

We present a semi-parametric approach to estimating item response functions (IRF) useful when the true IRF does not strictly follow commonly used functions. Our approach replaces the linear predictor of the generalized partial credit model with a monotonic polynomial. The model includes the regular generalized partial credit model at the lowest order polynomial. Our approach extends Liang's (2007) method for dichotomous item responses to the case of polytomous data. Furthermore, item parameter estimation is implemented with maximum marginal likelihood using the Bock-Aitkin EM algorithm, thereby facilitating multiple-group analyses useful in operational settings. Our approach is demonstrated on both educational and psychological data. We present simulation results comparing our approach to more standard IRF estimation approaches and other non-parametric and semi-parametric alternatives.

Keywords: Item response theory, semi-parametric models, monotonic polynomial, item response function

1 Introduction

Under many unidimensional item response theory (IRT) applications, it is common to assume that the underlying latent variable, θ, is normally distributed, and that item response functions (IRFs) follow one of many standard functions (e.g., Thissen & Steinberg, 1986). For example, it may be hypothesized that the relationship between the latent trait and probability of item response to dichotomous items has a logistic shape (i.e., 2-parameter logistic or 2PL), which provides a close approximation to the normal ogive model (Birnbaum, 1968). IRFs for ordered polytomous items may use the binary logistic function as "building blocks" (Thissen & Steinberg, 1986) to construct the graded response model (Samejima, 1969) or generalized partial credit model (GPC; Muraki, 1992).

In some cases, the assumption of a normally distributed latent trait or the assumption that the IRF follows one of these standard functions is violated. Proceeding with a standard approach may result in poor recovery of the true IRF and/or underlying latent trait estimates (e.g., Ramsay & Abrahamowicz, 1989; Liang, 2007). Methods have been proposed to estimate IRT models in the presence of each of these assumption violations, with it possible to model either a non-normally distributed latent trait *or* non-standard IRF, but not both simultaneously. For instance, if all non-standard IRFs tend to follow the same shape, researchers may suspect that a non-normally distributed latent trait is the culprit and choose to model the distribution using an empirical histogram approach (Mislevy, 1984; Woods, 2007a), Ramsay curves (Woods & Thissen, 2006; Woods, 2006, 2007b, 2008) or Davidian curves (Woods & Lin, 2008). Alternatively, non-standard IRFs can be modeled using techniques such as non-parametric methods (Ramsay, 1991; Rossi, Wang, & Ramsay, 2002; Samejima, 1977, 1979, 1984; Sijtsma, Debets, & Molenaar, 1990), semi-parametric methods (Liang, 2007; Ramsay & Winsberg, 1991), or by using Bayesian non-parametric estimation (Duncan & MacEachern, 2008, 2013; Miyazaki & Hoshino, 2009; Qin, 1998).

The purpose of this research is to present a semi-parametric technique for model-

ing non-standard IRFs that may have potential advantages over some existing methods. This work builds upon that of Liang (2007), who used a logistic function of a monotonic polynomial (MP) to estimate non-standard IRFs for dichotomous items. She developed an estimation technique that used provisional estimates of θ for each individual's latent trait (similar to Ramsay, 1991, 2000) in the complete-data log-likelihood. In this work, we present a multivariate logistic function of the MP that can model ordered polytomous items, which we refer to as the GPC-MP model since it reduces to the GPC model at the lowest order polynomial. When the number of categories is two, the model reduces further to the logistic function of a monotonic polynomial. When both of these conditions occur, the model reduces to the familiar 2PL. Furthermore, we utilize Bock-Aitkin (Bock & Aitkin, 1981) maximum marginal likelihood estimation using the expectation maximization algorithm (EM MML).

Use of the MP ensures that IRFs (or for polytomous data the basic building block functions) are monotonically increasing, which is not necessarily the case under other non-parametric models (e.g., Ramsay, 1991; Rossi et al., 2002).[1] Our approach does not require provisional θ estimates or information from an external test on the subject's ability (e.g., Liang, 2007; Ramsay, 1991; Samejima, 1977, 1979, 1984). Furthermore, use of EM MML facilitates mixing of different item types such as MP items and other item types within the same test, as well as multiple group analyses in which group means are compared or differential item functioning is investigated - features that are not always readily available under other approaches. Finally, to our knowledge this is the first approach to allow semi-parametric estimation of IRFs for polytomous items. Although several non-parametric techniques are available for polytomous items (Abrahamowicz & Ramsay, 1992; Mazza, Punzo, & McGuire, 2013; Ramsay, 2000; Santor, Ramsay, & Zuroff, 1994; Santor, Zuroff, Ramsay, Cervantes, & Palacios, 1995; van der Ark, 2007),

[1] We assume that monotonicity is desirable in many testing situations where a correct item will always indicate higher (or equal) ability for all regions of the latent trait. However, we do note that releasing constraints on monotonicity may be useful for probing for severe departures from monotonicity or when non-monotonicity is actually predicted.

the performance of such approaches are typically not compared to each other and, to our knowledge, do not include formal tests for differential item functioning.

The remainder of this manuscript is therefore organized as follows. Section 2 describes our parameterization of the MP and the GPC-MP for estimation of IRFs. Section 3 describes the details of our estimation procedure. In Section 4 we illustrate three potential applications of the GPC-MP model on real-world datasets. In Section 5 we present two simulation studies that examine test conditions under which the GPC-MP model performs better at recovering the true IRF versus standard approaches and a comparison of the GPC-MP model using EM MML against the alternatives of using Liang's (2007) estimation procedure or a non-parametric kernel smoothing approach (Ramsay, 2000; Mazza et al., 2013). Finally, Section 6 provides concluding remarks.

2 The Proposed Item Response Model

2.1 Monotonic Polynomial

The basic building block for the model we present is a monotonic polynomial and its first derivative of the form (e.g., see Liang, 2007; Heinzmann, 2005, 2008):

$$m(\theta|\xi, \mathbf{b}) = \xi + b_1\theta + b_2\theta^2 + \cdots + b_{2k+1}\theta^{2k+1} \tag{1}$$

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

Thus, ξ is an intercept parameter, and $\mathbf{b} = [b_1, \ldots, b_{2k+1}]$ and $\mathbf{a} = [a_0, \ldots, a_{2k}]$ are coefficients with $b_t = a_{t-1}/t$ for $t = 1, \ldots, 2k+1$. The degree of the polynomial is controlled by a user specified value for $k \in [0, \infty)$. For example, $k = 2$ would represent a 5th order polynomial. In our application, θ represents the latent trait. In one of its initial applications, the monotonic polynomial was used to estimate an unknown cumulative distribution function (CDF), with the monotonicity of the polynomial ensuring that the CDF be monotonically increasing (Heinzmann, 2005, 2008). The monotonicity requirement was established by ensuring that the polynomial was of an odd-order (by using

$2k+1$), and had a positive derivative $m'(\theta|\mathbf{a})$ for all θ. Elphinstone (1985) showed that the latter requirement could be accomplished by re-parameterizing the coefficients $\mathbf{a}$ in terms of λ, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k]$, and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_k]$:

$$m'(\theta|\lambda, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \begin{cases} \lambda \prod_{h=1}^{k} (1 - 2\alpha_h\theta + (\alpha_h^2 + \beta_h)\theta^2) & \text{if } k > 0 \\ \lambda & \text{if } k = 0 \end{cases} \tag{3}$$

and implementing the constraints that all $\boldsymbol{\beta} > 0$ and $\lambda > 0$. To later allow unconstrained estimation, we re-parameterize $\beta_1, \ldots, \beta_k = \exp(\tau_1), \ldots, \exp(\tau_k)$ and $\lambda = \exp(\omega)$, resulting in:

$$m'(\theta|\omega, \boldsymbol{\alpha}, \boldsymbol{\tau}) = \begin{cases} \exp(\omega) \prod_{h=1}^{k} (1 - 2\alpha_h\theta + (\alpha_h^2 + \exp(\tau_h))\theta^2) & \text{if } k > 0 \\ \exp(\omega) & \text{if } k = 0 \end{cases} \tag{4}$$

Liang (2007) noted that the coefficients for $m'(\theta|\lambda, \boldsymbol{\alpha}, \boldsymbol{\beta})$, $\mathbf{a}_k = [\ a_0 \ \cdots \ a_{2k}\ ]'$, for a polynomial of degree $2k+1$ could be represented in recursive form as $\mathbf{a}_k = \mathbf{T}_k\mathbf{a}_{k-1} = \mathbf{T}_k\mathbf{T}_{k-1}\cdots\mathbf{T}_2\mathbf{T}_1\lambda$ with any given matrix $\mathbf{T}_k$ containing parameters only corresponding to α_k and β_k, and a special case being that $\mathbf{a}_0 = \lambda$. Under our re-parameterization, each matrix $\mathbf{T}_k$ contains only parameters α_k and τ_k and the expression for the coefficients $\mathbf{a}_k$ becomes:

$$\mathbf{a}_k = \mathbf{T}_k\mathbf{a}_{k-1} = \mathbf{T}_k\mathbf{T}_{k-1}\cdots\mathbf{T}_2\mathbf{T}_1 \exp(\omega) \tag{5}$$

The matrices $\mathbf{T}_k$ have dimensions $(2k+1) \times (2k-1)$ and are of the form:

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

The advantage of the recursive form for $\mathbf{a}_k$, as described by Liang (2007), is that it is relatively easy to code a computer program to construct the polynomial $m(\theta|\xi, \omega, \boldsymbol{\alpha}, \boldsymbol{\tau})$ and its derivative $m'(\theta|\omega, \boldsymbol{\alpha}, \boldsymbol{\tau})$ on the fly up to an arbitrary order. We start at $k = 0$ to compute $\mathbf{a}_0$. We then move on to $k = 1$ by $\mathbf{a}_1 = \mathbf{T}_1\mathbf{a}_0$, and so on, working our way up until the desired k. The coefficients $\mathbf{b}_k$ necessary for constructing the polynomial can then be computed from $\mathbf{a}_k$. If any derivatives of $\mathbf{a}_k$ with respect to the parameters $\boldsymbol{\tau}$ and $\boldsymbol{\alpha}$ are required, we only need to differentiate and save the corresponding matrix $\mathbf{T}$ that contains such parameters. For instance, $\frac{\partial \mathbf{a}_3}{\partial \alpha_1} = \mathbf{T}_3\mathbf{T}_2\frac{\partial \mathbf{T}_1}{\partial \alpha_1}\exp(\omega)$, since α_1 is only contained in $\mathbf{T}_1$.

2.2 A Monotonic Polynomial Generalized Partial Credit Model

Suppose we have $i = 1, 2, \ldots, N$ independently sampled individuals respond to $j = 1, 2, \ldots, n$ items, and $c = 0, 1, \ldots C_j - 1$ coding the response categories for item j. When items are dichotomous (i.e., $C_j = 2$), the two-parameter logistic (2PL) model can be used to represent the probability that individual i will endorse category 1 on item j:

$$P(1|\theta_i, \delta_j, \gamma_j) = \frac{1}{1 + \exp(-(\delta_j + \gamma_j\theta_i))} \tag{6}$$

where θ_i is the person's latent trait, γ_j is a slope parameter, and δ_j is the intercept.

With polytomous items ($C_j > 2$), the generalized partial credit model can be used to represent the probability that respondent i's response to item j is category c (Muraki, 1992), and can be parameterized in slope/intercept form as:

$$P(c|\theta_i, \boldsymbol{\delta}_j, \gamma_j) = \frac{\exp\left[\sum_{v=0}^{c}(\delta_{jv} + \gamma_j\theta_i)\right]}{\sum_{u=0}^{C_j-1}\exp\left[\sum_{v=0}^{u}(\delta_{jv} + \gamma_j\theta_i)\right]} \tag{7}$$

with $\boldsymbol{\delta}_j = [\delta_{j0}, \ldots, \delta_{C_j-1}]$ essentially representing more than one intercept parameter. This divide-by-total model is a constrained version of Bock's nominal response model (e.g., Thissen, Cai, & Bock, 2010), in which there is a single slope parameter for each item (or alternatively that the slope parameter for all categories on a single item are constrained equal). This model reduces to the 2PL when $C_j = 2$.

It is easy to see how a monotonic polynomial can be substituted for the linear predictors of the above item models for a more flexible modeling approach. Liang (2007) was the first to describe the logistic function of a monotonic polynomial as simply the 2PL model, but with the linear predictor replaced by the monotonic polynomial for item j:

$$P(1|\theta_i, \xi_j, \omega_j, \boldsymbol{\alpha}_j, \boldsymbol{\tau}_j) = \frac{1}{1 + \exp(-m_j(\theta_i, \xi_j, \omega_j, \boldsymbol{\alpha}_j, \boldsymbol{\tau}_j))} \tag{8}$$

The fact that $m_j(\theta_i, \xi_j, \omega_j, \boldsymbol{\alpha}_j, \boldsymbol{\tau}_j)$ is monotonically increasing ensures that the item response function for this item is also monotonically increasing. At the lowest-order polynomial (i.e., $k = 0$), this model reduces to the 2PL with the additional constraint that the slope parameter is positive. At $k \geq 1$, one or more "bends" in the item response function may occur, resulting in flexible IRF shapes.

A generalization of the above model is a GPC model that makes use of the monotonic

polynomial in its linear predictor, which we abbreviate GPC-MP, constructed as:

$$P(c|\theta_i, \boldsymbol{\xi}_j, \omega_j, \boldsymbol{\alpha}_j, \boldsymbol{\tau}_j) = \frac{\exp\left[\sum_{v=0}^{c}(\xi_{jv} + m_j^*(\theta_i, \omega_j, \boldsymbol{\alpha}_j, \boldsymbol{\tau}_j))\right]}{\sum_{u=0}^{C_j-1}\exp\left[\sum_{v=0}^{u}(\xi_{jv} + m_j^*(\theta_i, \omega_j, \boldsymbol{\alpha}_j, \boldsymbol{\tau}_j))\right]} \tag{9}$$

where $m_j^*(\theta_i, \omega_j, \boldsymbol{\alpha}_j, \boldsymbol{\tau}_j) = b_1\theta_i + \cdots + b_{2k+1}\theta_i^{2k+1}$ is the monotonic polynomial without the intercept term. The set of C_j multinomial logits in Equation (9) share the same MP term but have different intercepts. This model reduces to the GPC model when $k = 0$, making m_j^* a linear function of θ. When $C_j = 2$, this model reduces to the logistic function of a monotonic polynomial. When both $k = 0$ and $C_j = 2$, we have a 2PL model. Furthermore, it can be shown that the conditional response probability of category c given the response is in category c or c' is a logistic function of a monotonic polynomial. For the GPC model the intercepts are overparameterized and hence one strategy employed by Muraki (1992) for identification is to constrain the first intercept ξ_{j0} to zero. We take this same approach to identify the GPC-MP model.

To illustrate the different shapes that the GPC-MP model can take, several example IRFs and accompanying parameters appear in Figure 1 and Table 1. These IRFs were obtained by fitting models with all items modeled as $k = 1, 2$ or 3 to data later described in Sections 4.1 and 4.3 for dichotomous items (Examples 1-3) and 5-category items (Examples 4-6). In general, the order of the polynomial will determine the number of possible "bends"' in the IRFs, though some departures from the 2PL and GPC are more noticeable than others. In examining dichotomous items (left column of Figure 1), the departures from a 2PL can be mild (Example 1), appear as flat regions in locations where we might otherwise expect a lower or upper asymptote before dropping down to 0 or up to 1 (Example 2), and can be small flat regions or mild kinks within the same IRF (Example 3). For polytomous items (right column of Figure 1), wherever there is a non-standard bend in the IRF, it will appear in the same region of θ across response functions for all categories. Note also that $\exp(\omega)$ is no longer directly interpretable as

an overall slope of the item as it is under the 2PL and GPC model, but is interpretable as the slope of a tangent line when $\theta = 0$. For instance, Examples 4 and 6 have very small values ($\exp(\omega) = .17$ and .03) where there is a flat region across all IRFs at $\theta = 0$, but Example 5 does not share this property. The remaining $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ parameters determine curvature changes of the IRF and their interpretation is nontrivial.

3 Parameter Estimation

In this section we describe two different estimation approaches for the GPC-MP model, beginning with complete data estimation as used in the surrogate-based approach used by Liang (2007) and Ramsay (1991). The complete data models are then used to derived the EM MML estimation method. Additional details relevant to our implementation of the GPC-MP model are also discussed.

3.1 Complete Data Likelihood and Surrogate-Based Estimation

Let y_{ij} represent the response from individual i to item j. We denote

$$f(y_{ij}|\theta_i, \boldsymbol{\eta}_j) = \prod_{c=0}^{C_j-1} P(c|\theta_i, \boldsymbol{\eta}_j)^{\chi_c(y_{ij})} \tag{10}$$

as the conditional density of y_{ij} given the latent trait θ_i. All parameters for item j are contained in a vector $\boldsymbol{\eta}_j$, and $\chi_c(y_{ij})$ is an indicator function that is equal to 1 if and only if $y_{ij} = c$, and 0 otherwise. Note that $P(c|\theta_i, \boldsymbol{\eta}_j)$ is a general expression that may be any of the item response functions presented in Section 2. Under the local independence assumption (Lord & Novick, 1968), the conditional density of person i's response vector $\mathbf{y}_i$ is then:

$$f(\mathbf{y}_i|\theta_i, \boldsymbol{\eta}) = \prod_{j=1}^{n} f(y_{ij}|\theta_i, \boldsymbol{\eta}_j). \tag{11}$$

The complete data likelihood is then:

$$L(\boldsymbol{\eta}, \mu, \sigma^2|\boldsymbol{\theta}, \mathbf{Y}) = \left[\prod_{i=1}^{N} f(\mathbf{y}_i|\theta_i, \boldsymbol{\eta})\right] \left[\prod_{i=1}^{N} \phi(\theta_i|\mu, \sigma^2)\right] \tag{12}$$

where the bracket on the right-hand side includes the population distribution of θ, which in this case is assumed to be normal. For typical single-sample calibration, μ and σ^2 are fixed for identifiability of the model, yet we retain these parameters in the following notation to later facilitate description of multiple group estimation. The log-likelihood can be partitioned as

$$l(\boldsymbol{\eta}, \mu, \sigma^2 | \boldsymbol{\theta}, \mathbf{Y}) = \log L(\boldsymbol{\eta} | \boldsymbol{\theta}, \mathbf{Y}) + \log L(\mu, \sigma^2 | \boldsymbol{\theta}) \tag{13}$$

where the part pertaining to item parameters corresponds to:

$$l(\boldsymbol{\eta} | \boldsymbol{\theta}, \mathbf{Y}) = \log L(\boldsymbol{\eta} | \boldsymbol{\theta}, \mathbf{Y}) = \sum_{j=1}^{n} \left[\sum_{i=1}^{N} \sum_{c=0}^{C_j - 1} \chi_c(y_{ij}) \log P(c | \theta_i, \boldsymbol{\eta}_j) \right]. \tag{14}$$

Note that Equation (14) is a set of n independent multinomial logistic regression log-likelihoods. If the latent traits were known we could proceed to estimate item parameters by differentiating $l(\boldsymbol{\eta} | \boldsymbol{\theta}, \mathbf{Y})$ with respect to each model parameter, and solve the resulting system of likelihood equations through iterative methods (e.g., Newton-Raphson).

Ignoring constants, the part that corresponds to the population distribution mean and variance parameters can be written as:

$$l(\mu, \sigma^2 | \boldsymbol{\theta}) = \log L(\mu, \sigma^2 | \boldsymbol{\theta}) \propto -\frac{N}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (\theta_i - \mu)^2. \tag{15}$$

For this complete data model, linear sufficient statistics exist. They are the sum and sum-of-squares of the latent trait values: $\sum_{i=1}^{N} \theta_i$ and $\sum_{i=1}^{N} \theta_i^2$.

Prior to estimation of model parameters, Liang (2007) and Ramsay (1991, 2000) compute an approximation to the latent traits, $\boldsymbol{\theta}$, which are then assumed known in subsequent estimation of model parameters. Although using an altogether different approach to estimating IRFs, in Ramsay (1991, 2000) respondents are ranked using weighted sum scores, $\sum_{j=1}^{n} \sum_{c=0}^{C_j - 1} \chi_c(y_{ij}) w_{jc}$, with weights determined by $w_{jc} = logit(P_{jc}^{(75)}) -$

$logit(P_{jc}^{(25)})$, where $P_{jc}^{(75)}$ and $P_{jc}^{(25)}$ represent the proportion of respondents in the upper and lower 25% of the sum score distribution that responded c for item j. Any ties among ranks across respondents with different response vectors may be broken randomly. The ranks are then transformed onto the quantiles of some distribution, usually standard normal, and used as provisional estimates of θ. Liang (2007) instead used the respondents' scores on the first principal component (of the covariance matrix of the items) as the ranking basis and normalized the resulting scores. Liang (2007) then used these scores as approximate θ estimates for use in the optimization of the complete data log-likelihood. Following Liang (2007), we refer to this as a *surrogate-based* (SB) approach.

3.2 EM MML Estimation

Our preferred approach is to estimate the GPC-MP model with maximum marginal likelihood using the Bock-Aitkin EM algorithm (Bock & Aitkin, 1981). We assume a population distribution, such as a normal distribution, $\phi(\theta_i|\mu, \sigma^2)$. The marginal (observed data) distribution of a response vector is obtained by marginalization over the unobserved latent traits (Bock & Lieberman, 1970):

$$L(\boldsymbol{\eta}, \mu, \sigma^2|\mathbf{y}_i) = f(\mathbf{y}_i|\boldsymbol{\eta}, \mu, \sigma^2) = \int f(\mathbf{y}_i|\theta_i, \boldsymbol{\eta})\phi(\theta_i|\mu, \sigma^2)d\theta_i \tag{16}$$

And the marginal log-likelihood is given by $l(\boldsymbol{\eta}, \mu, \sigma^2|\mathbf{Y}) = \sum_{i=1}^{N} \log L(\boldsymbol{\eta}, \mu, \sigma^2|\mathbf{y}_i)$.

Integration over θ must be conducted numerically. For simplicity we employ a Q-point rectangular quadrature rule where the quadrature points $X_1, X_2, \ldots X_Q$ are equally spaced (e.g., between -6 and 6 on θ). Serving as quadrature weights are normalized population distribution ordinates, i.e., $W_q = \phi(X_q|\mu, \sigma^2) / \sum_{q=1}^{Q} \phi(X_q|\mu, \sigma^2)$.

The Bock-Aitkin EM algorithm alternates between two steps. In the E-step, item parameters are assumed known and θ is treated as missing data. The conditional expectation of the complete data log-likelihood is taken with respect to the posterior $f(\theta|\mathbf{y}_i, \boldsymbol{\eta}, \mu, \sigma^2)$. In the M-step, the conditional expected log-likelihood is optimized,

resulting in updated parameter estimates. The two steps alternate until only a small change in parameter estimates and/or log-likelihoods (e.g., .001) occurs across successive iterations. We refer the reader to the Appendices for additional details.

3.3 A Model with Multiple Groups

Assuming $g = 1, 2, \ldots, G$ groups, the marginal log-likelihood for the combined multiple group model is the sum of the within-group marginal log-likelihoods:

$$l(\boldsymbol{\eta}, \boldsymbol{\mu}, \sigma^2 | \mathbf{Y}) = \sum_{g=1}^{G} l(\boldsymbol{\eta}_g, \mu_g, \sigma_g^2 | \mathbf{Y}_g). \tag{17}$$

As will be illustrated later in empirical examples, two possible uses of multiple group estimation are 1) the ability to fix some (or all) item parameters equal across groups to link the scale and estimate whether a studied group's latent distribution has a different mean (and/or variance) versus a reference group with a fixed mean and variance, and 2) test for differential item functioning by estimating whether the order of a polynomial for a particular item is the same (or different) across groups and whether item parameters are equal across groups. Both of these investigations require implementation of equality constraints on some item parameters across groups. For this purpose we estimated all item parameters simultaneously and used Lagrange multipliers (e.g., Bertsekas, 1996).

Nested models under this approach can be compared using a likelihood ratio test. Similarly, GPC-MP items with higher-order polynomials require the addition of unconstrained α and τ parameters and can be compared versus models with lower-order polynomials using such a test (Liang, 2007).

3.4 Additional Estimation Details

In initial tests of the GPC-MP model, some item parameters encountered an ill-conditioned likelihood, and thus caused estimation problems and a rank deficient Hessian. We imposed diffuse priors on a subset of item parameters to aid in estimation (e.g., Cai, Yang, & Hansen, 2011; Woods & Thissen, 2006), which technically results in

Bayesian estimation of the posterior maximum for model parameters. However, such priors may be considered "soft constraints" that serve merely to stabilize estimation as is commonly employed with the three-parameter logistic model (Cai et al., 2011); their use usually results in trivial changes in the estimated IRF and log-likelihood. Based on some preliminary tests of the model, we used diffuse priors of $\pi(\tau) \sim N(-1, 500)$ and $\pi(\alpha) \sim N(0, 500)$. The choice of a prior mean of 0 for α and a negative value for τ roughly correspond to values that would cause the GPC-MP model to reduce to the GPC model; although τ would need to be $-\infty$ for this to occur, small negative values appeared to have a similar effect as more extreme negative values.

All programming and simulations were conducted in R (R Core Team, 2012) by the first author. Newton-Raphson with a simple backtracking algorithm was implemented for maximization of the complete-data and M-step log-likelihoods: the step size was successively halved (until 1e-6) if the updated parameter estimates did not lead to an increase in log-likelihood. Hessians were conditioned by ridging.

Conceptually, the approach used by Liang (2007) is similar to computing a provisional estimate of $\boldsymbol{\theta}$ and then performing a single M-step to estimate the model parameters. It should be noted that there were some additional minor differences between our version of the surrogate-based approach and Liang's actual implementation. Specifically, Liang did not re-parameterize $\boldsymbol{\beta}$ and λ from the monotonic polynomial and instead opted to use constrained estimation to enforce all $\boldsymbol{\beta} > 0$ and $\lambda > 0$. Furthermore, Liang did not utilize diffuse priors on any model parameters and did not ridge the Hessians, but instead used the negative information matrix.

3.5 Model Selection

Since the GPC-MP model can take a different degree of polynomial, $2k + 1$, for each item, some selection process may be necessary to avoid blindly picking k. To automate this process, we primarily experimented with using Akaike's Information Criterion (AIC), which is defined as $AIC = -2 \log L + 2(\#\text{ of parameters})$, where $\log L$ is the

log-likelihood. In the surrogate-based approach, $\log L$ is based on the complete-data log-likelihood, whereas under EM MML, $\log L$ is the marginal log-likelihood. In a subset of examples and simulations, we implemented an AIC step-wise approach in which we started with a baseline model where all items were $k = 0$. We looped over all items, setting each to $k = 1$ while keeping all other items the same as in the previous step, calculated the change in AIC and chose to increase the degree of polynomial for the item that improved AIC the most. This process repeated until no additional items modeled as $k = 1$ could improve AIC. The resulting model was used for a starting point in looping over all items as $k = 2$, and so on. For the surrogate-based approach, this step-wise model is efficient to compute since θ estimates are fixed and we would only need to loop over all items once; the change in AIC is not dependent upon whether other items are modeled with a higher degree polynomial. For EM MML, this process is slower as provisional expected θ estimates at each quadrature point may change and modeling one item as $k = 1$, for example, will affect whether modeling a different item to a higher degree polynomial will improve AIC. Thus, we also experimented with a more crude but efficient approach in which models with *all* items were estimated as $k = 0, 1$ or 2 in simulations (up to $k = 3$ in empirical examples) and AIC was used to select among these models. We refer to these selected models as *AIC step-wise* and *AIC selected*, respectively.

4 Empirical Examples

The feasibility of the GPC-MP using EM MML is now demonstrated using examples from educational and psychological assessments. We show that for empirical datasets model fit may improve by including some items as higher-order polynomials. The potential applications of GPC-MP using EM MML estimation to test mean differences in the latent trait across groups and in differential item functioning are also illustrated.

4.1 Probing for Non-Standard IRFs

We analyzed responses from 32 items from the Program for International Student Assessment (PISA) Reading Book 8 from the year 2000 using a sample of 3000 participants

as described by Cai (2010). Data were in the form of dichotomous scores (incorrect, correct) for most items, and partial credit (0,1,2) awarded for three additional items. Column 1 of Table 5 shows the number of categories for each item. The purpose of this illustration was merely to probe for non-standard IRFs in an educational testing situation that included a mix of dichotomous and polytomous items. We estimated five models with the first four where all items were modeled as $k = 0, 1, 2, 3$ and a final model where the degree of the polynomial for each item was selected by AIC in a step-wise fashion as described in the previous section.

As shown in Table 2, the best fitting model was the step-wise model, followed by the $k = 1$ model. The GPC-MP model resulted in modest improvement over standard GPC and 2PL (i.e., $k = 0$) items, and over half of the AIC step-wise items were modeled with higher order polynomials: 8 items modeled as 5th-order polynomials and 13 items as 3rd-order polynomials (see Table 5). Estimated item response functions under the AIC step-wise model appear in Figure 2. Some $k = 1$ items had merely a small kink in the IRF (e.g., Items 2, 7, 26, 30, and 32) whereas $k = 2$ modeled items had greater non-trivial departures from a logistic IRF (e.g., Items 1, 4, 6, 9, 15, 20, 23, and 24).

4.2 Estimation of Group Mean Differences

The calibration data from the Patient Reported Outcomes Measurement Information System (PROMIS®) smoking module (Hansen et al., in press; Shadel, Edelen, & Tucker, 2011) included responses from 4,201 daily smokers (28-30 days smoked out of the past 30 days) and 1,183 non-daily smokers (<28 days smoked out of the past 30 days) to items divided among 6 different constructs related to smoking behavior: nicotine dependence, hedonic benefits, coping benefits, social benefits, psychosocial risks, and health risks. Data collection followed a randomized block design using 26 overlapping forms to reduce respondent burden. Items had five possible ordered response options.

For illustration purposes, we analyzed responses to only the hedonic benefits item bank, which included 16 items for daily smokers and 17 items for non-daily smokers,

15 of which were common to both groups. Thus, the goals of this illustration were to 1) Model the 15 common items using the GPC-MP model and AIC to select the order of the polynomial for each item; and 2) Estimate the mean difference in daily/non-daily smokers on the underlying hedonic benefits latent trait.

To accomplish these goals, the AIC step-wise approach described in Section 3 up to $k = 3$ was used while the daily smoker mean and variance of hedonic benefits were fixed to a standard normal distribution and the non-daily mean and variance was freely estimated. Since the original calibration analyses suggested that the 15 common items differed negligibly in their item parameters across groups, parameters for these items were constrained equal across groups. The items not shared by both groups were fitted with the GPC model.

As shown in Table 5 and Figure 3, AIC selection of k for each item suggested several departures from the standard GPC model, including seven items with a 7th-order polynomial ($k = 3$). The final model also indicated that non-daily smokers had a lower mean on hedonic benefits ($\hat{\mu} = -.42$) than the daily smokers ($\hat{\mu} = 0$), $\chi^2(1) = 76.72$, $p < .001$ (see also Table 3).

4.3 Differential Item Functioning

The GPC-MP model using EM MML estimation also may have utility in situations where the order of polynomial for a particular item may differ across groups, and may additionally be used to test whether higher order polynomials have the same item parameters across groups. For this illustration, we used the social benefits domain from the PROMIS® smoking module, which included 12 items for daily and non-daily smokers, 9 of which were common to both group's item bank. Although initial calibration results indicated few items with substantial differential item functioning as indicated by comparing the expected score of the IRFs across groups (Hansen et al., in press), for illustration purposes we selected only two anchors (items 5 and 7) modeled as standard GPC items. The remaining 7 items were studied one-by-one using the GPC-MP model

while freely estimating the mean and variance of the reference group's latent trait.

One possible procedure is to first determine the order of polynomial for each group by fitting a series of models whereby the studied item is $k = 0, 1, 2, 3$ for a single group, and use likelihood ratio tests to determine whether any of the higher order polynomials presented significant improvement over lower-order polynomials.[2] Although this does not explore all possible combinations of k for the item across groups, as with use of AIC this process should provide a reasonable guess as the to order of polynomial for each group as supported by the data.

The above procedure resulted in only item 1 being modeled as the same order polynomial ($k = 1$) across groups (see also Table 5). Constraining all 7 parameters for this item equal across groups resulted in significant worsening of model fit, $\chi^2(7) = 30.30$, $p < .001$, suggesting differential item functioning for this item. Constraining only the slope and bend parameters, ω, α, and τ equal across groups did not result in worse fit, $\chi^2(3) = 2.13$, $p = .55$, analogous to failing to provide evidence for non-uniform differential item functioning. Additionally constraining intercept parameters, $\boldsymbol{\xi}$, equal across groups resulted in significant worsening of model fit, $\chi^2(4) = 28.17$, $p < .001$, analogous to providing evidence for uniform differential item functioning.

5 Simulation Studies

Two simulation studies were performed. The goal of the first was to test item response function recovery of the GPC-MP model using EM MML under a variety of conditions and against the standard 2PL and GPC model. The second simulation was intended as a more focused comparison of GPC-MP using EM MML versus the alternative IRF estimation procedures introduced by Liang (2007) and Ramsay (1991).

The main outcome of these simulations was IRF recovery as measured by a generalization of Root Integrated Mean Square Error (RIMSE; e.g., Liang, 2007; Ramsay, 1991)

[2] We implemented this approach using a conventional $p < .05$ threshold for hypothesis testing though note that Benjamini-Hochberg adjustment is sometimes used in practice to control the false discovery rate in differential item functioning situations (e.g., see Thissen, Steinberg, & Kuang, 2002).

for polytomous item j:

$$\text{RIMSE}_j = \left(\frac{\sum_{q=1}^{Q} (\hat{ES}_j(\theta_q) - ES_j(\theta_q))^2 \phi(\theta_q)}{\sum_{q=1}^{Q} \phi(\theta_q)} \right)^{1/2} \times 100 \tag{18}$$

where $ES_j(\theta_q) = \sum_{c=0}^{C_j-1} c \cdot P(c|\theta_q, \boldsymbol{\eta}_j)/(C_j - 1)$ is the expected score for item j at a given level of θ (divided by $C_j - 1$ to place dichotomous and polytomous items on the same scale), the sum is across quadrature points, and ϕ is the density function for a standard normal distribution. The secondary outcome was latent trait recovery as measured by a similar index:

$$\text{RMSE}_\theta = \left(N^{-1} \sum_{i=1}^{N} (\hat{\theta}_i - \theta_i)^2 \right)^{1/2} \times 100 \tag{19}$$

where θ_i is the actual latent trait for subject i and $\hat{\theta}_i$ is the estimated trait.[3] Unless noted otherwise, trait estimates were calculated using the *expected a posteriori* (EAP) scoring method (Bock & Mislevy, 1982). Conceptually, RIMSE_j measures the agreement of the estimated IRF with that of the true IRF with more weight given to estimates near the center of the latent trait distribution. To provide a single index of IRF (and trait) recovery for each cell in the simulation design, RIMSE_j (and RMSE_θ) was further aggregated across items (or respondents) for each replication and again across all replications in a single cell.

5.1 Study 1: IRF Recovery Versus Standard Approaches

5.1.1 Data Generation

In assessing the GPC-MP model's IRF recovery, we generated unidimensional data crossing the following data generation conditions: 1) Number of categories per item (2 and 5), 2) True polynomial order ($k = 1$ and $k = 2$), 3) Number of test items (10 and 20) and 4) Sample size (500 and 3000). One-hundred data sets per cell were generated. In all cases, the latent traits were drawn from a standard normal distribution. For half

[3] We thank an anonymous reviewer for suggesting this index for trait recovery. Results for latent trait recovery using RIMSE_θ as in Liang (2007) are available from the authors upon request.

of the test items, the true item parameters were generated randomly across replications. For dichotomous items, $\xi \sim \text{unif}(-1,1)$, $\exp(\omega) \sim \text{unif}(.3,2.5)$, $\alpha \sim \text{unif}(-1,1)$, and $\exp(\tau) \sim \text{unif}(0,1)$, which are similar to values used in simulations by Liang (2007). For polytomous (5 category) items, $\xi_0 = 0$, $\xi_1 \sim \text{unif}(1,2.5)$, $\xi_2 \sim \text{unif}(-1,1)$, $\xi_3 \sim \text{unif}(-1,1)$, $\xi_4 \sim \text{unif}(-2.5,-1)$ and $\exp(\omega) \sim \text{unif}(.3,1.5)$.

The true parameters for the remaining items were the same across replications. Dichotomous item parameters for $k = 1$ and $k = 2$ were drawn from initial calibration results for five different items estimated using a subset of the PISA dataset. Polytomous item parameters were drawn from five items of the PROMIS® social benefits domain as calibrated on daily smokers only. For the 20 item cell, parameters from five items were repeated to obtain parameters for 10 items. Selection of these item parameters was based on which items had $k = 1$ or $k = 2$ models that improved AIC and a visual inspection to include extreme-looking items and a variety of IRF shapes.

5.1.2 Fitted Models

In each cell of the design, EM MML was used to fit a series of GPC-MP models. This included three models that modeled all items using the same degree polynomial from $k = 0,1,2$, and a fourth AIC selected model that merely picked among the all $k = 0,1$ or 2 models. In a subset of conditions (10 items), we included use of the AIC step-wise model starting at $k = 0$ and working up to $k = 2$. All models were estimated using 49 equally spaced quadrature points between -5 and 5, and estimation terminated upon changes in item parameters less than .001.

5.1.3 Results

RIMSE results for items appear in Figure 4; lower RIMSE values are better. To facilitate comparisons among fitted models, for any given data generation combination, the fitted models were ranked (1 = best, 5 = worst) with ties allowed and shading assigned to these ranks (white = better, gray = worse). As our key interest is whether higher-order polynomials have better IRF recovery than lower-order polynomials and whether

AIC selection performed well, this shading strategy emphasizes differences across fitted models, but not across data generation conditions.

Even when the true model was a higher-order polynomial (e.g., $k = 2$), IRF recovery was not always best by fitting the true model to every item, but this trend depended on the number of categories per item. For dichotomous items, IRF recovery was often better with a lower-order polynomial (e.g., $k = 0$ or $k = 1$). This was especially the case with few respondents and items. For instance, IRF recovery at $N = 500$ and 10 items was best with $k = 0$ when the true $k = 1$ (RIMSE $= 5.14$) and $k = 1$ when the true $k = 2$ (RIMSE $= 5.43$). At $N = 3000$ and 20 items, however, IRF recovery was best when the true model was fit to the data (RIMSE $= 1.75$ at $k = 1$ and RIMSE $= 2.59$ at $k = 2$). With polytomous items, IRF recovery appeared to be at or near best when fitting the true model, regardless of number of items and sample size. For example, IRF recovery was always first or second best among fitted models when all $k = 1$ items were modeled as $k = 1$, and the same for when $k = 2$ items were modeled as such.

With respect to the utility of AIC in selecting the degree of each items' polynomial, the step-wise model tended to be third best (6/8 cells) or better in the conditions in which it appeared. Although this performance appears modest at best, note also that the AIC step-wise model never had RIMSE values more than .3 behind the top performing model, whereas the worst-performing models could have RIMSE values that were 3 or more behind. With the exception of only two cells in the design in which it was less than .1 behind the second place model (both at $N = 500$ and 10 items), the cruder AIC selected model was always either first or second in IRF recovery. The AIC selected model performed exceptionally well in conditions with more items and respondents, in which it was always first or tied for first with RIMSEs ranging from .96 to 2.59.

Recovery of latent traits as measured by RMSE_θ nearly mirrored recovery of IRFs (see Figure 5) in that latent trait recovery tended to be best in cases when IRF recovery was best. For example, under 10 items and $N = 3000$, the ranking of the methods was iden-

tical to IRF recovery for all cells but three. Although the ranking of methods departed slightly more from IRF results under other conditions, it was still true that latent trait recovery tended to be better for higher-order polynomial models under conditions of more information (polytomous items, more items, more subjects). Results for the AIC selected model were similar in that it was ranked first or second in all cases but one at $N = 500$ and 10 items. The AIC step-wise model also performed decently in terms of latent trait recovery, being ranked third in 5/8 cases, and first under two other conditions. Overall these results suggest that use of the GPC-MP model may be best with a high degree of information, but also that the use of AIC can aid in choosing the degree of the polynomial for each item when the true data generating model is not known.

5.2 Study 2: Comparison Versus Alternative Estimation Approaches

5.2.1 Data Generation

In order to compare the GPC-MP model to alternatives, we crossed the following data generation conditions: 1) Number of categories per item (2 and 3), and 2) Sample size (500 and 3000). In each cell of the data generation design, one-hundred datasets with twenty items each were generated. Data generation differed from that of the previous study in that the true IRFs were *not* from GPC-MP items, but were instead generated from a mixture of the cumulative distribution function (CDF) of normal variables so that our approach is not implicitly favored.

Half of the IRFs under dichotomous items were generated following Liang (2007) who used two normal CDFs, $p_1 N(\mu_1, \sigma_1^2) + p_2 N(\mu_2, \sigma_2^2)$, with $p_1 \sim \text{unif}(.3, .7)$, $p_2 = 1 - p_1$, $\mu_1 \sim N(-1.5, .1^2)$, $\mu_2 \sim N(1, .1^2)$, $\sigma_1 \sim N(1, .1^2)$, $\sigma_2 \sim N(.4, .1^2)$. In our subjective opinion, the resulting IRFs under such data generating values were often very similar and not very extreme. Thus, the remaining dichotomous items were a mixture of three CDFs with p_1 and $p_2 \sim \text{unif}(.1, .4)$, $p_3 = 1 - p_1 - p_2$, $\mu_1 \sim N(-1.5, .4^2)$, $\mu_2 \sim N(1.5, .4^2)$, $\mu_3 \sim N(0, .4^2)$, and σ_1, σ_2, and σ_3 independently drawn from $N(.4, .1^2)$. Polytomous (3 category) items were generated using a mix of 2 or 3 (half of the items for each)

normal CDFs pieced together in an analogous fashion to a graded response model, i.e., $P(y_{ij} \geq c) = p_1 N(\mu_1 + \iota_c, \sigma_1^2) + p_2 N(\mu_2 + \iota_c, \sigma_2^2)$, with $P(y_{ij} \geq 0) = 1$ and $\iota_c > \iota_{c-1}$ since $\iota_1 \sim \text{unif}(-1.5, -.5)$ and $\iota_2 \sim \text{unif}(.5, 1.5)$. All μ and σ values were drawn using the same distributions as dichotomous items.

5.2.2 Fitted Models

The main goal was to compare the following approaches: 1) GPC-MP using EM MML; 2) GPC-MP using the surrogate-based approach; and 3) The kernel smoothing approach of Ramsay (1991). To this end, GPC-MP models were estimated in the same manner as in the previous simulation study (all $k = 0, 1, 2$ models, the AIC step-wise model, and AIC selection of $k = 0, 1, 2$ models). In addition, these models were estimated using EM MML and the surrogate-based approach. Finally, the ksIRT function from the *KernSmoothIRT* (Mazza et al., 2013) package in R was used to estimate IRFs under the kernel smoothing approach using 51 equally spaced points between -3 and 3 on θ. Note also that ksIRT generates only maximum likelihood-based latent trait estimates; we report the ML estimates but also added a normal prior for computation of EAP estimates for this method.

Theoretical expectations about the expected pattern of results are difficult to form. Note that our implementation of the EM and SB approaches are the same except that the latent traits, θ, are treated as missing data under EM but approximate estimates are used in the complete data likelihood under SB. To our knowledge, the performance of such initial latent trait estimates under SB has not been studied versus the EM MML solution. Had further iterations of the SB approach been employed alternating between trait and item parameter estimates, it may resemble the joint maximum likelihood approach. The KS approach also uses an altogether different approach to initial θ estimates and a non-parametric approach to estimating IRFs. Finally, the data generating IRFs are not derived from any of the estimation approaches we employed, and the use of a graded-type (i.e., differencing) model to piece together IRFs for polytomous items may present an

additional challenge for the GPC-MP items (which are divide-by-total models).

5.2.3 Results

IRF recovery for the various modeling approaches appears in Figure 6 using the same shading strategy as in Study 1. At $N = 500$, the KS approach emerged as the top performing method by .08 and .23 over the second best method (under dichotomous and polytomous items, respectively), with the remaining results somewhat mixed. In examining all $k = 0, 1, 2$ for both SB and EM MML, the all $k = 1$ models performed well under dichotomous items at that sample size, suggesting that higher-order polynomials were appropriate for the IRFs, but the $k = 0$ model performed best under polytomous items. In comparing the utility of AIC, for dichotomous items, EM MML ($k = 1$ and AIC selected) and SB with the AIC selected model were tied for second with dichotomous items, followed by the AIC step-wise model for EM MML. For polytomous items, AIC step-wise and selected for EM MML performed second and fourth best, respectively, suggesting that EM MML combined with some form of AIC selection is viable. At $N = 3000$, the advantage of higher-order polynomials and EM MML over SB emerged more clearly. With all $k = 2$, EM MML occupied the first and second place (tied) spots for dichotomous and polytomous items, respectively, with the EM MML AIC step-wise and selected models occupying places between second and fourth. The best competing method was the KS approach, which was .72 worse than the EM step-wise model under dichotomous items, but .11 better under polytomous items.

Recovery of latent traits presented a similar pattern to that of IRF recovery with a few exceptions (see Figure 7). At $N = 500$, KS with EAP scoring performed better by .02 under dichotomous and by 2.49 under polytomous items over the second best method. Under dichotomous items at $N = 500$, SB and EM MML with higher order polynomials (anything but $k = 0$) or AIC step-wise/selection performed well, but with SB having a slight advantage. With polytomous items at $N = 500$, $k = 0$ models performed better than those with higher-order polynomials and AIC selection/step-wise

models performed better for EM MML than for SB. As with IRF recovery, at $N = 3000$ the advantage of the GPC-MP model with EM MML began to emerge along with a clearer pattern of results. The $k = 2$, AIC selected, and AIC step-wise models for EM MML occupied the top two spots (with ties) along with SB with AIC selection. Under polytomous items, these same models occupied spots behind the KS method with EAP and ML scoring. ML estimates using the KS approach, the default estimates provided by KernSmoothIRT (Mazza et al., 2013), tended to perform worse than other methods and was always ranked last with the exception of its second place performance under $N = 3000$ with polytomous items.

6 Discussion

We have presented a monotonic polynomial generalized partial credit model that subsumes the 2PL and GPC models and is estimated using EM MML. This model has the potential to probe for and model non-standard item response functions and has potential applications in multi-group analyses. In addition, our simulation studies demonstrate that the use of higher order polynomials can result in better IRF and latent trait recovery when the true IRF does not follow a 2PL or GPC shape. Furthermore, the GPC-MP item model using EM MML estimation tended to perform better than the SB approach of Liang (2007) for IRF recovery and was nearly on par with the IRF recovery of the Kernel smoothing approach of Ramsay (1991).

That said, we note some important limitations and potential directions for future research. First, as shown in Study 1, use of a lower-order polynomial can sometimes lead to better IRF recovery than use of a higher-order polynomial even when the true IRF follows the higher-order polynomial. We speculate that when insufficient information exists (e.g., few items, subjects), that the use of higher-order polynomials is prone to fitting noise in a similar way to how the kernel smoothing approach may result in non-monotonic IRFs (especially in the tails) even when the true IRF is monotonically increasing. Although not explicitly manipulated in our studies, we would also suspect

that more discriminating items may aid in providing more information necessary for fitting the GPC-MP.

To enhance the usability of the GPC-MP, a mechanism may be required for either probing for higher-order polynomials and/or testing whether a particular item is better modeled as a higher-order polynomial. In our simulations, AIC demonstrated promise in selecting the order of polynomial. AIC selected and step-wise models were not always the best out of fitted models, but were never far behind and are clearly preferable to blindly picking the order of the polynomial. Although our choice of AIC is somewhat arbitrary, Liang (2007) experimented with use of AIC, BIC and likelihood ratio tests, finding that these approaches performed similarly. We suggest that any of these approaches remain as a potential direction for additional research, and add that examination of item fit statistics (Orlando & Thissen, 2000) also remains a possibility. We speculate that BIC would result in fewer items being modeled as higher-order polynomials due to its bias towards parsimony, and would thus allow researchers to model more items as 2PL or GPC items and reduce computational time if a BIC step-wise model were implemented.

Under our simulations, better IRF recovery for the GPC-MP model typically translated into better latent trait recovery. One exception to this pattern was at $N = 500$ with dichotomous items for Study 2, where better IRF recovery using EM MML did not necessarily lead to better trait recovery than the SB approach. Since these two approaches both used the GPC-MP model, without additional research it is difficult to determine whether there is a bias/efficiency trade-off to latent trait recovery under these two estimation approaches or if another explanation exists. To our knowledge, the SB approach used by Liang (2007) has not been extensively studied in previous research in comparison to EM MML, nor has the use of provisional θ estimates by Ramsay (1991). We note that the current studies were primarily geared towards a comparison of IRF recovery and not specifically designed with an investigation of trait recovery in mind. Thus, as the KS approach tended to perform near the top in terms of IRF recovery under Study

2, we also do not know if this is due to kernel smoothing of IRFs, surrogate θ estimates that are better than Liang's use of the subjects' scores on the first principal component, or whether piecing together CDFs as in the graded response model instead of a divide-by-total manner put the GPC-MP at a disadvantage for polytomous items. Nonetheless, even though true items were *not* generated by the GPC-MP model, the GPC-MP model using AIC step-wise or selection was never far behind the KS approach, and sometimes outperformed it.

Several possible variations of the GPC-MP model and comparisons with other approaches also remain for future research. For instance, should researchers wish to estimate a GPC function of a polynomial that has a negative relationship with the latent trait *or* without the monotonicity requirement, constraints on λ *or* both λ and $\boldsymbol{\beta}$ could be released, respectively, by not reparameterizing. While we do not rule out the possibility of constructing GPC functions of other types of polynomials (e.g., splines), such an approach will likely not offer the same level of parsimony as GPC-MP, and it is unclear whether this would have any advantage over existing approaches (e.g., monotone splines; Ramsay & Winsberg, 1991). Finally, our approach has not yet been compared with other promising, though computationally intensive Bayesian non-parametric approaches (e.g., Duncan & MacEachern, 2008; Miyazaki & Hoshino, 2009).[4]

In conclusion, use of the GPC-MP using EM MML has potential advantages over the KS and SB approaches. In the realm of multiple-group analysis, EM MML provides a more natural way of estimating group differences and testing of differential item functioning across groups. Whereas the KS approach models all items non-parametrically, the GPC-MP model has the potential to allow researchers to model some or most items on a test as standard 2PL and GPC items. Thus, use of the GPC-MP allows greater parsimony, potentially easier interpretation of item parameters in a test, and formal testing of whether a higher-order polynomial fits better than a lower-order polynomial.

[4]See Liang (2007) for a comparison of a logistic function of the monotonic polynomial using surrogate-based estimation with the methods of Duncan and MacEachern (2008) and Qin (1998).

Appendix A. Further Derivations of EM MML Estimation

Recall that the complete data likelihood is $L(\boldsymbol{\eta}, \mu, \sigma^2|\mathbf{y}_i, \theta_i) = f(\mathbf{y}_i|\theta_i, \boldsymbol{\eta})\phi(\theta_i|\mu, \sigma^2)$ for individual i. Thus, an individual's contribution to the marginal likelihood can be approximated to arbitrary precision as

$$\bar{L}_i(\boldsymbol{\eta}, \mu, \sigma^2) = \sum_{q=1}^{Q} f(\mathbf{y}_i|X_q, \boldsymbol{\eta})W_q.$$

Using quadrature, the height of the complete data likelihood at quadrature point X_q can be represented as

$$L(\boldsymbol{\eta}, \mu, \sigma^2|\mathbf{y}_i, X_q) = \prod_{j=1}^{n}\prod_{c=0}^{C_j-1} P(c|X_q, \boldsymbol{\eta}_j)^{\chi_c(y_{ij})}W_q.$$

This suggests that the ordinate of the posterior $f(\theta|\mathbf{y}_i, \boldsymbol{\eta}, \mu, \sigma^2)$ at quadrature point X_q can be approximated by

$$f(X_q|\mathbf{y}_i, \boldsymbol{\eta}, \mu, \sigma^2) = \frac{f(\mathbf{y}_i, X_q|\boldsymbol{\eta}, \mu, \sigma^2)}{f(\mathbf{y}_i|\boldsymbol{\eta}, \mu, \sigma^2)} \approx \frac{L(\boldsymbol{\eta}, \mu, \sigma^2|\mathbf{y}_i, X_q)}{\bar{L}_i} = \bar{P}_i(X_q). \tag{20}$$

For the item parameter part, the conditional expectation of $\log L(\boldsymbol{\eta}|\theta_i, \mathbf{y}_i)$ is

$$E_i(\boldsymbol{\eta}|\boldsymbol{\eta}_*, \mu_*, \sigma_*^2) = \int \left[\sum_{j=1}^{n}\sum_{c=0}^{C_j-1} \chi_c(y_{ij}) \log P(c|\theta_i, \boldsymbol{\eta}_j)\right] f(\theta_i|\mathbf{y}_i, \boldsymbol{\eta}_*, \mu_*, \sigma_*^2)d\theta_i,$$

where $\boldsymbol{\eta}_*$, μ_*, and σ_*^2 are the current/provisional parameter estimates. Using Equation (20), this conditional expectation may be approximated by quadrature as

$$E_i(\boldsymbol{\eta}|\boldsymbol{\eta}_*, \mu_*, \sigma_*^2) \approx \sum_{q=1}^{Q} \left[\sum_{j=1}^{n}\sum_{c=0}^{C_j-1} \chi_c(y_{ij}) \log P(c|X_q, \boldsymbol{\eta}_j)\right] \bar{P}_i(X_q).$$

Summing over all N individuals and rearranging terms in the summation, the condi-

tional expectation of $\log L(\boldsymbol{\eta}|\boldsymbol{\theta}, \mathbf{Y})$ from Equation (14) is

$$\begin{aligned} E(\boldsymbol{\eta}|\boldsymbol{\eta}_*, \mu_*, \sigma_*^2) &= \sum_{i=1}^{N} E_i(\boldsymbol{\eta}|\boldsymbol{\eta}_*, \mu_*, \sigma_*^2) \\ &\approx \sum_{i=1}^{N}\sum_{q=1}^{Q}\sum_{j=1}^{n}\sum_{c=0}^{C_j-1} \chi_c(y_{ij}) \log P(c|X_q, \boldsymbol{\eta}_j)\bar{P}_i(X_q) \\ &= \sum_{j=1}^{n}\sum_{q=1}^{Q}\sum_{c=0}^{C_j-1} \bar{r}_{jqc} \log P(c|X_q, \boldsymbol{\eta}_j), \end{aligned} \tag{21}$$

where $\bar{r}_{jqc} = \sum_{i=1}^{N} \chi_c(y_{ij})\bar{P}_i(X_q)$ is the conditional expected frequencies for item j, category c, at quadrature point q. Item parameter estimates are updated in the M-step by treating the expected frequencies as weights and maximizing $E(\boldsymbol{\eta}|\boldsymbol{\eta}_*, \mu_*, \sigma_*^2)$.

For distributional parameters for the latent traits, μ and σ^2, if not fixed in a multiple group analysis, can be also be estimated by calculating the mean and variance of the expected counts (e.g., see Baker & Kim, 2004). This is possible because the conditional expectations of the linear sufficient statistics may also be approximated via quadrature and the M-step is closed-form. For example, the conditional expectation of $\sum_{i=1}^{N} \theta_i$ is

$$E\left(\sum_{i=1}^{N} \theta_i|\boldsymbol{\eta}_*, \mu_*, \sigma_*^2\right) \approx \sum_{i=1}^{N}\sum_{q=1}^{Q} X_q\bar{P}_i(X_q) = \sum_{q=1}^{Q}\left(\sum_{i=1}^{N}\bar{P}_i(X_q)\right)X_q = \sum_{q=1}^{Q}\bar{r}_q X_q,$$

where $r_q = \sum_{i=1}^{N}\bar{P}_i(X_q)$ is the conditional expected frequencies at quadrature point q. An updated latent variable mean estimate is therefore $N^{-1}\sum_{q=1}^{Q}\bar{r}_q X_q$. For the variance parameter, the conditional expectation of $\sum_{i=1}^{N}\theta_i^2$ is

$$E\left(\sum_{i=1}^{N} \theta_i^2|\boldsymbol{\eta}_*, \mu_*, \sigma_*^2\right) \approx \sum_{i=1}^{N}\sum_{q=1}^{Q} X_q^2\bar{P}_i(X_q) = \sum_{q=1}^{Q}\left(\sum_{i=1}^{N}\bar{P}_i(X_q)\right)X_q^2 = \sum_{q=1}^{Q}\bar{r}_q X_q^2.$$

An updated variance estimate is found by $N^{-1}\sum_{q=1}^{Q}\bar{r}_q X_q^2 - \left(N^{-1}\sum_{q=1}^{Q}\bar{r}_q X_q\right)^2$.

Appendix B. Complete Data Derivatives for the GPC-MP Model

The complete data log-likelihood for a single GPC-MP item j is:

$$l_j = \sum_{i=1}^{N} \sum_{c=0}^{C_j-1} \chi_c(y_{ij}) \log P(c|\theta_i, \boldsymbol{\xi}_j, \omega_j, \boldsymbol{\alpha}_j, \boldsymbol{\tau}_j).$$

where $P(c|\theta_i, \boldsymbol{\xi}_j, \omega_j, \boldsymbol{\alpha}_j, \boldsymbol{\tau}_j)$ is the item response function for item j under the GPC-MP model. Differentiating with respect to a typical parameter, η_t, leads to:

$$\frac{\partial l_j}{\partial \eta_t} = \sum_{i=1}^{N} \left[\sum_{c=0}^{C_j-1} \chi_c(y_{ij}) \left(\sum_{v=0}^{c} \left(\frac{\partial m_{ij}^*}{\partial \eta_t} + \frac{\partial \xi_{jv}}{\partial \eta_t} \right) - \sum_{u=0}^{C_j-1} \sum_{v=0}^{u} P_{iju} \left(\frac{\partial m_{ij}^*}{\partial \eta_t} + \frac{\partial \xi_{jv}}{\partial \eta_t} \right) \right) \right],$$

where $m_{ij}^* = m_j^*(\theta_i, \omega_j, \boldsymbol{\alpha}_j, \boldsymbol{\tau}_j)$ is short-hand for the monotonic polynomial without the intercept parameter, and $P_{iju} = P(u|\theta_i, \boldsymbol{\xi}_j, \omega_j, \boldsymbol{\alpha}_j, \boldsymbol{\tau}_j)$ is short-hand for person i's probability of responding to category u on item j under the GPC-MP model. Of course, $\frac{\partial m_{ij}^*}{\partial \xi_{jv}} = 0$, and $\frac{\partial \xi_{jv}}{\partial \eta_t}$ is 1 when differentiating with respect to ξ_{jv} and 0 otherwise. The derivatives of m_{ij}^* for the parameters ω_j, α_{js} and τ_{js} are simply the following and can be substituted into the above equation:

$$\begin{aligned}
\frac{\partial m_{ij}^*}{\partial \omega_j} &= \frac{\partial m_{ij}^*}{\partial \mathbf{a}} \mathbf{T}_k \mathbf{T}_{k-1} \cdots \mathbf{T}_2 \mathbf{T}_1 \exp(\omega_j) \\
\frac{\partial m_{ij}^*}{\partial \alpha_{js}} &= \frac{\partial m_{ij}^*}{\partial \mathbf{a}} \mathbf{T}_k \mathbf{T}_{k-1} \cdots \frac{\partial}{\partial \alpha_{js}} \mathbf{T}_s \cdots \mathbf{T}_2 \mathbf{T}_1 \exp(\omega_j) \\
\frac{\partial m_{ij}^*}{\partial \tau_{js}} &= \frac{\partial m_{ij}^*}{\partial \mathbf{a}} \mathbf{T}_k \mathbf{T}_{k-1} \cdots \frac{\partial}{\partial \tau_{js}} \mathbf{T}_s \cdots \mathbf{T}_2 \mathbf{T}_1 \exp(\omega_j),
\end{aligned}$$

in which $s = 1, 2, \ldots, k$, the $\mathbf{T}$ matrices are specific to item j, and $\frac{\partial m_{ij}^*}{\partial \mathbf{a}}$ is the vector:

$$\frac{\partial m_{ij}^*}{\partial \mathbf{a}} = \begin{bmatrix} \theta_i & \frac{1}{2}\theta_i^2 & \frac{1}{3}\theta_i^3 & \cdots & \frac{1}{2k+1}\theta_i^{2k+1} \end{bmatrix}.$$

The derivatives of the matrices $\mathbf{T}$ with respect to α_{js} and τ_{js} have the form:

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

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

For computing the Hessian, we used the following cross-product of gradients approximation, in which derivative vectors are computed for each individual and their outer-products are accumulated (e.g., Bock & Lieberman, 1970): $\sum_{i=1}^{N}\left(\frac{\partial l_j(\boldsymbol{\eta}_j|\theta_i,y_{ij})}{\partial \boldsymbol{\eta}_j}\right)\left(\frac{\partial l_j(\boldsymbol{\eta}_j|\theta_i,y_{ij})}{\partial \boldsymbol{\eta}_j}\right)'$. The above derivatives can be easily adapted for the M-step computations in EM MML estimation by summing over the Q quadrature points instead of over the N individuals, and by treating the expected frequencies $\bar{r}_{jqc}$ at each quadrature point as weights.

References

Abrahamowicz, M., & Ramsay, J. O. (1992). Multicategorical spline model for item response theory. *Psychometrika*, *57*(1), 5-27.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Bertsekas, D. P. (1996). *Contrained optimization and Lagrange multiplier methods*. Belmont, MA: Athena Scientific.

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., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika*, *35*, 179–197.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612.

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.

Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. G., & Edelen, M. O. (in press). Methodology for developing and evaluating the PROMIS smoking item banks. *Nicotine & Tobacco Research*.

Heinzmann, D. (2005). *A filtered polynomial approach to density estimation*. Unpublished master's thesis, Institute of Mathematics, University of Zurich.

Heinzmann, D. (2008). A filtered polynomial approach to density estimation. *Computational Statistics*, *23*, 343-360.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mazza, A., Punzo, A., & McGuire, B. (2013). *KernSmoothIRT: Nonparametric Item Response Theory. R Package Version 5.0.* Retrieved from `http://CRAN.R-project.org/package=KernSmoothIRT`

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*(3), 359-381.

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.

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

R Core Team. (2012). *R: A language and environment for statistical computing.* Vienna, Austria. Retrieved from `http://www.R-project.org` (ISBN 3-900051-07-0)

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

Ramsay, J. O. (2000). TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data [Computer software].

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

Ramsay, J. O., & Winsberg, S. (1991). Maximum marginal likelihood estimation for semiparametric item analysis. *Psychometrika*, *56*(3), 365-379.

Rossi, N., Wang, X., & Ramsay, J. O. (2002). Nonparametric item response estimates with the EM algorithm. *Journal of Educational and Behavioral Statistics*, *27*(3), 291-317.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs*, *17*.

Samejima, F. (1977). A method of estimating item characteristic functions using the maximum likelihood estimate of ability. *Psychometrika*, *42*.

Samejima, F. (1979). *A new family of models for the multiple choice item* (Tech. Rep. No. 79-4). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1984). *A plausibility function of Iowa Vocabulary Test items estimated by the simple sum procedure of the conditional P.D.F. approach* (Tech. Rep. No. 84-1). Knoxville: University of Tennessee, Department of Psychology.

Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment*, *6*(3), 255-270.

Santor, D. A., Zuroff, D. C., Ramsay, J. O., Cervantes, P., & Palacios, J. (1995). Examining scale discriminability in the BDI and CES-D as a function of depressive severity. *Psychological Assessment*, *7*(2), 131-139.

Shadel, W. G., Edelen, M., & Tucker, J. S. (2011). A unified framework for smoking assessment: The PROMIS smoking initiative. *Nicotine & Tobacco Research*, *13*(5), 399-400.

Sijtsma, K., Debets, P., & Molenaar, I. (1990). Mokken scale analysis for polychotomous items: theory, a computer program and an empirical application. *Quality & Quantity*, *24*, 173-188.

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Developments and applications* (p. 43-75). New York, NY: Taylor & Francis.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567–577.

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.

van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*, 1-19.

Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods*, *11*, 253-270.

Woods, C. M. (2007a). Empirical histograms in item response theory with ordinal data. *Educational and Psychological Measurement*, *67*, 73-87.

Woods, C. M. (2007b). Ramsay curve IRT for Likert-type data. *Applied Psychological Measurement*, *31*(3), 195-212.

Woods, C. M. (2008). Ramsay curve item response theory for the three-parameter item response theory model. *Applied Psychological Measurement*, *36*(6), 447-465.

Woods, C. M., & Lin, N. (2008). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement*, *33*(2), 102-117.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, *71*, 281–301.

Table 1: Item Parameters from Example IRFs

Parameter	Example 1	Example 2	Example 3	Example 4	Example 5	Example 6
ξ_1	0.37	0.71	-0.90	0.14	3.48	-0.21
ξ_2				0.18	2.50	-0.24
ξ_3				-1.06	-0.25	-1.20
ξ_4				-1.52	-1.64	-0.93
ω	-0.11	0.69	0.54	-1.76	1.02	-3.41
α_1	0.24	-0.50	-0.73	3.80	0.89	-0.53
α_2		0.52	0.81		-0.74	0.52
α_3			0.36			8.20
τ_1	-0.21	-8.48	-6.65	-1.79	-8.70	-9.92
τ_2		-3.32	-1.96		-8.99	-5.91
τ_3			-8.26			-1.50

Note. Examples 1 to 3 are dichotomous items with $k = 1, 2$ and 3, respectively. Examples 4 to 6 are 5-category items with $k = 1, 2$ and 3, respectively.

Table 2: Model Overview for PISA data

	All $k = 0$ (GPC)	All $k = 1$	All $k = 2$	All $k = 3$	AIC step-wise
# Parameters	67	131	195	259	125
$-2 \log L$	106007.9	105761.1	105644.8	105597.6	105704.6
AIC	106141.9	<u>106023.1</u>	106034.8	106115.6	<u>105954.6</u>

Note. Best two AIC values are underlined.

Table 3: Model Overview for PROMIS® Hedonic Benefits data

	All $k = 0$ (GPC)	AIC step-wise	Constrained Means
# Parameters	92	158	157
$-2 \log L$	124715.3	124296.2	124373.0
AIC	124899.3	124612.2	124687.0

Note. In all models common item parameters were constrained equal across daily and non-daily smokers. Group means were constrained equal in the "constrained means" model, but the non-daily mean was free in the other models.

Table 4: DIF models studying item 1 for PROMIS® Social Benefits data

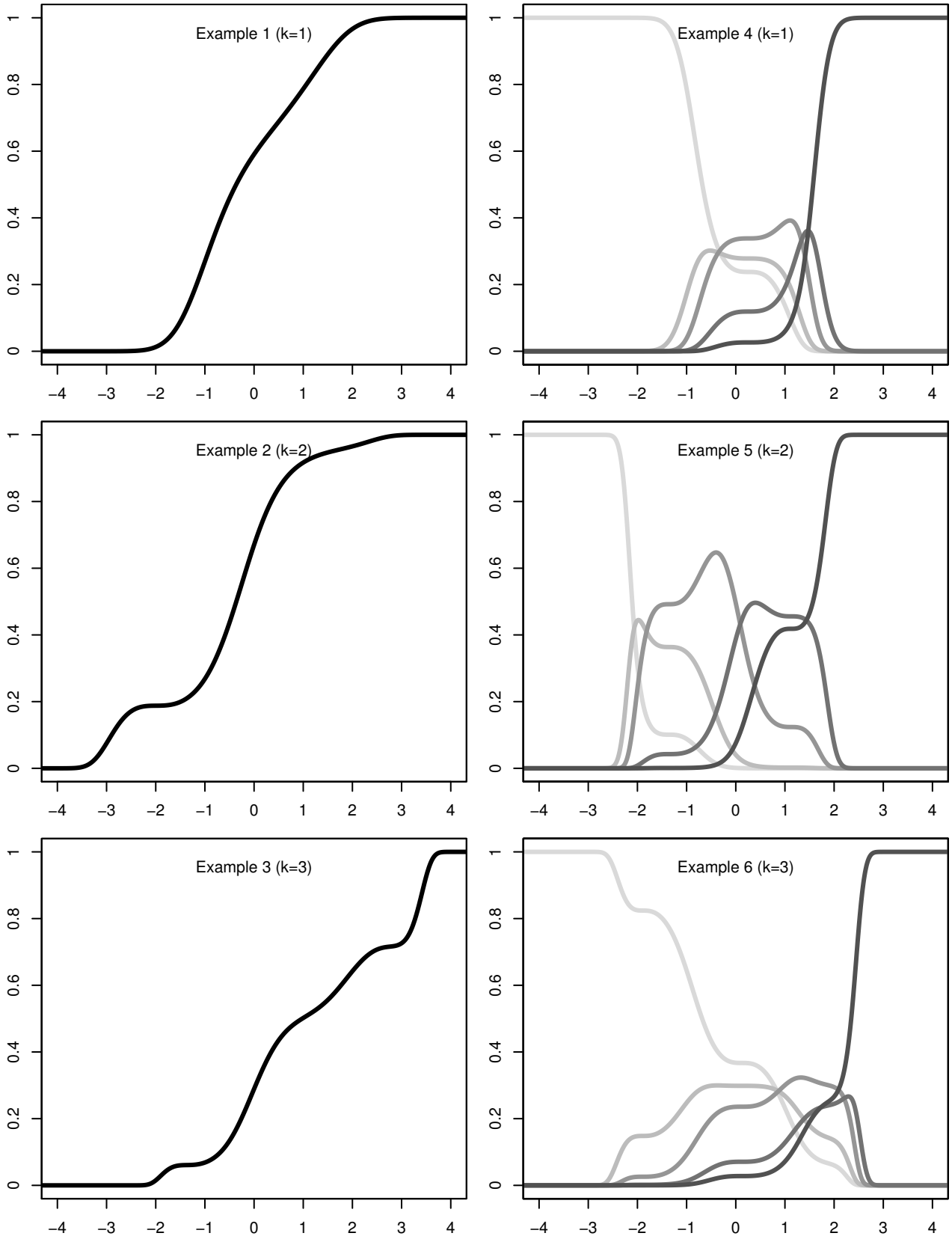
	Unconstrained	Equal ω, α, and τ	Equal ω, α, τ, and $\boldsymbol{\xi}$
# Parameters	116	113	109
$-2\log L$	88640.2	88642.4	88670.5
AIC	88872.2	88868.4	88888.5

Table 5: Item Overview for PISA and PROMIS® data

	PISA		Hedonic Benefits		Social Benefits		
Item	C_j	Selected k	C_j	Selected k	C_j	Selected k_1	Selected k_2
1	2	2	5	2	5	1	1
2	2	1	5	2	5	1	0
3	2	1	5	1	5	1	0
4	2	2	5	1	5	0	1
5	2	1	5	1	5	0	0
6	2	2	5	3	5	1	2
7	3	1	5	1	5	0	0
8	2	0	5	3	5	1	0
9	2	2	5	3	5	0	1
10	2	0	5	3			
11	2	0	5	3			
12	3	1	5	3			
13	3	0	5	3			
14	2	0	5	2			
15	2	2	5	2			
16	2	0					
17	2	0					
18	2	1					
19	2	0					
20	2	2					
21	2	1					
22	2	1					
23	2	2					
24	2	2					
25	2	0					
26	2	1					
27	2	0					
28	2	1					
29	2	1					
30	2	1					
31	2	0					
32	2	1					

Note. For PROMIS® data (hedonic benefits and social benefits), only items common to both daily and non-daily smokers are presented. k_1 refers to daily smokers and k_2 refers to non-daily smokers. Items 5 and 7 for social benefits served as anchors. C_j is number of categories per item. k was selected using AIC for PISA and hedonic benefits, and likelihood ratio tests for social benefits.

Figure 1: Example IRFs



Note. Response curves for higher categories are darker.

Figure 2: IRFs for PISA AIC Stepwise Model

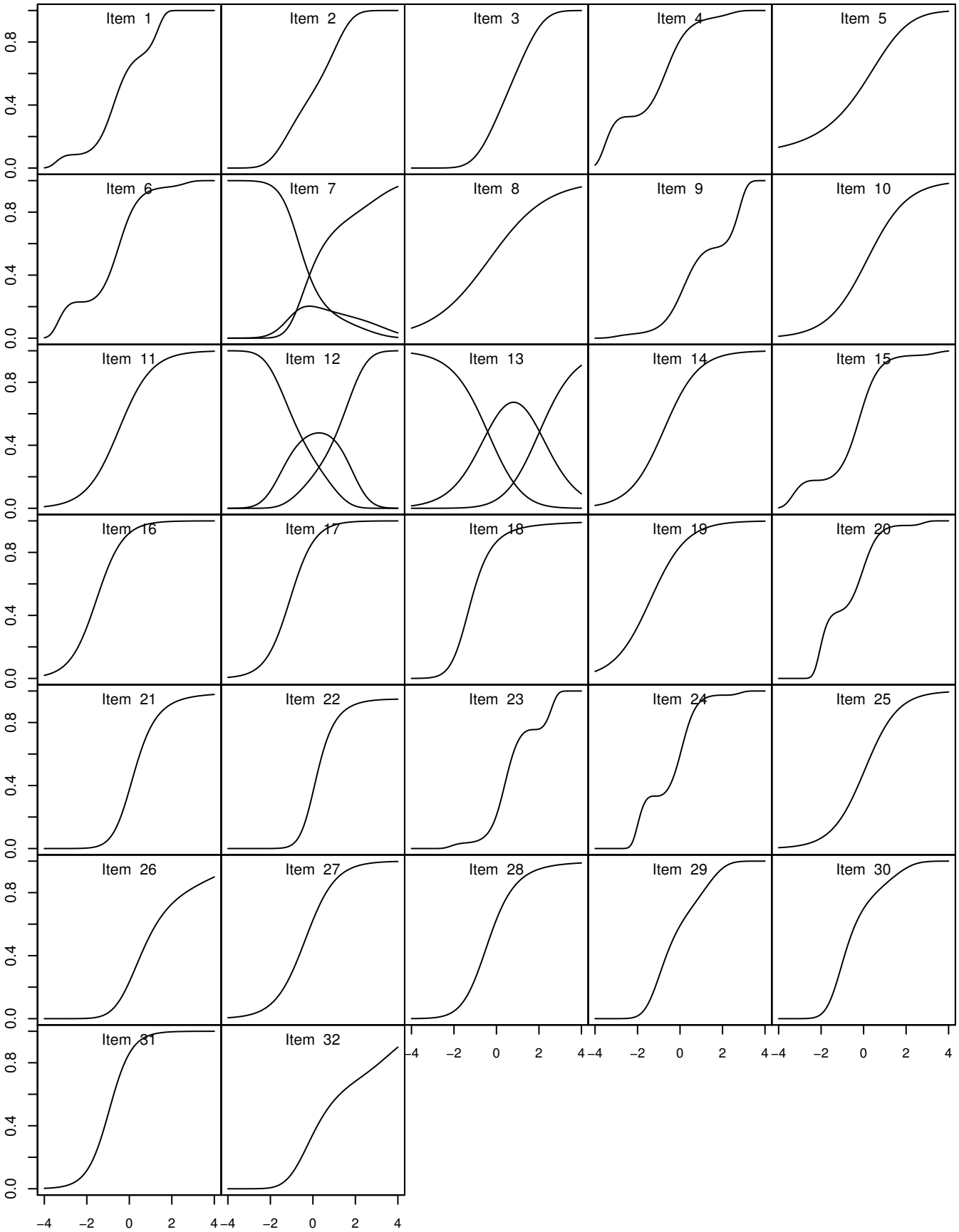
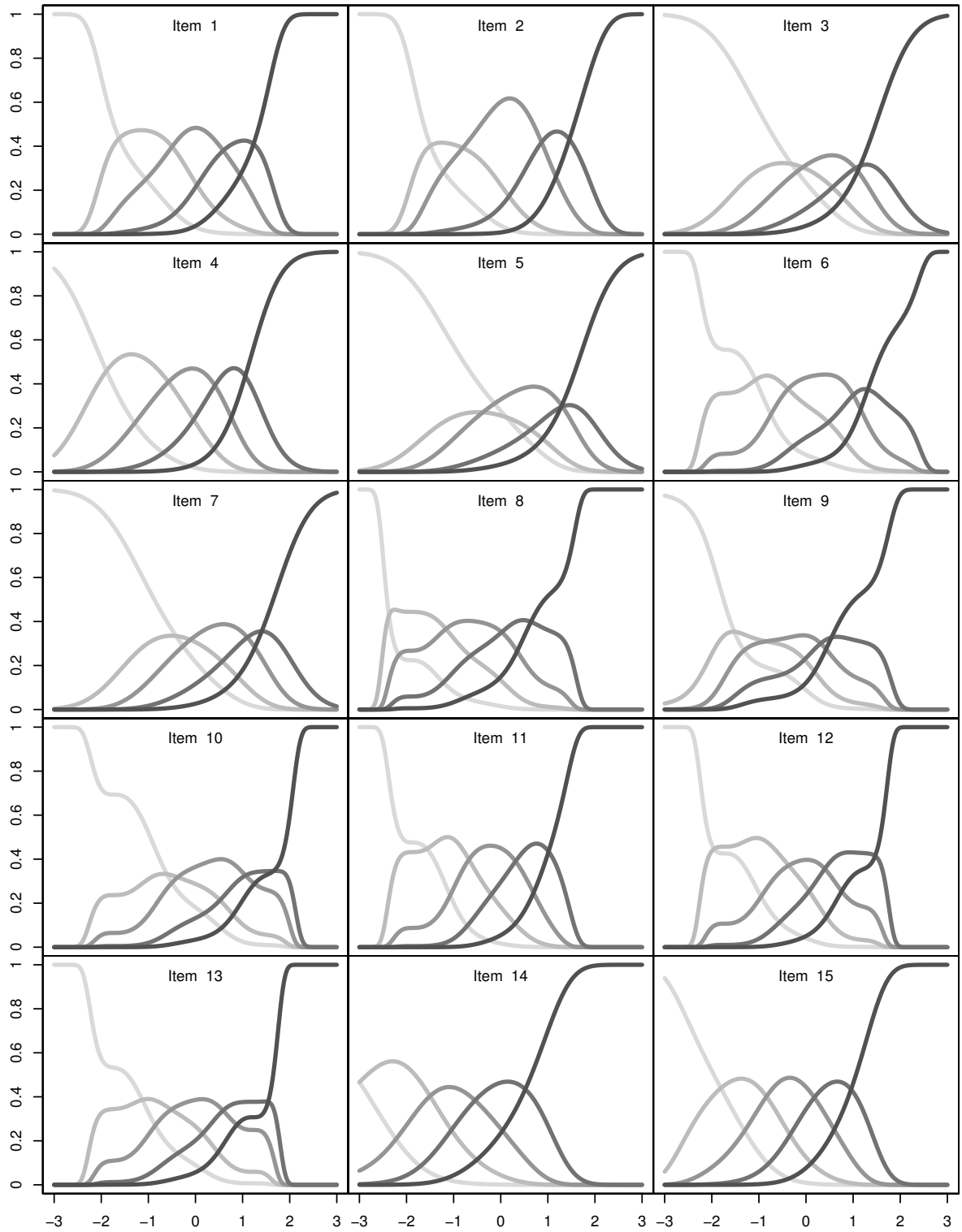
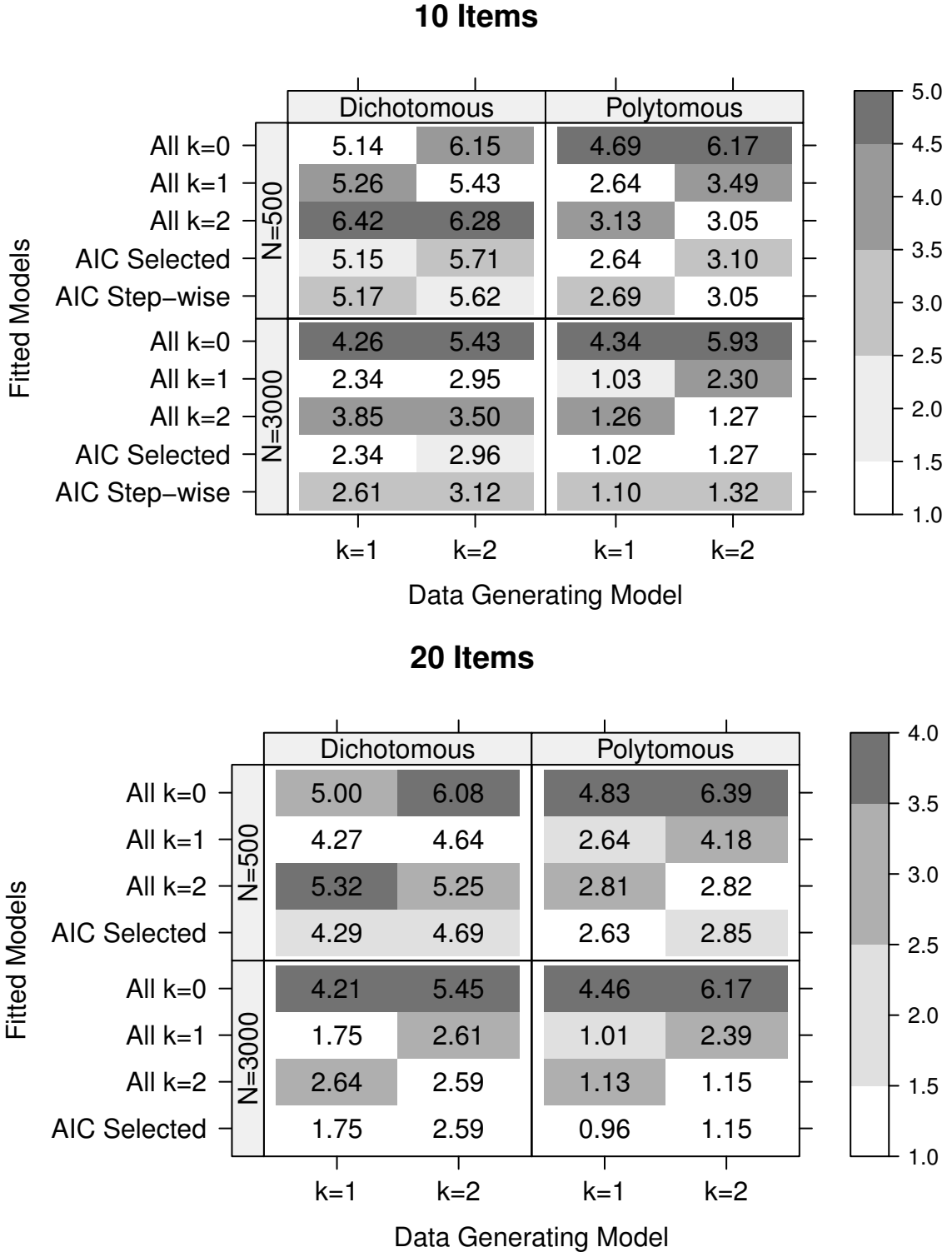


Figure 3: IRFs for PROMIS® Hedonic Benefits AIC Stepwise Model



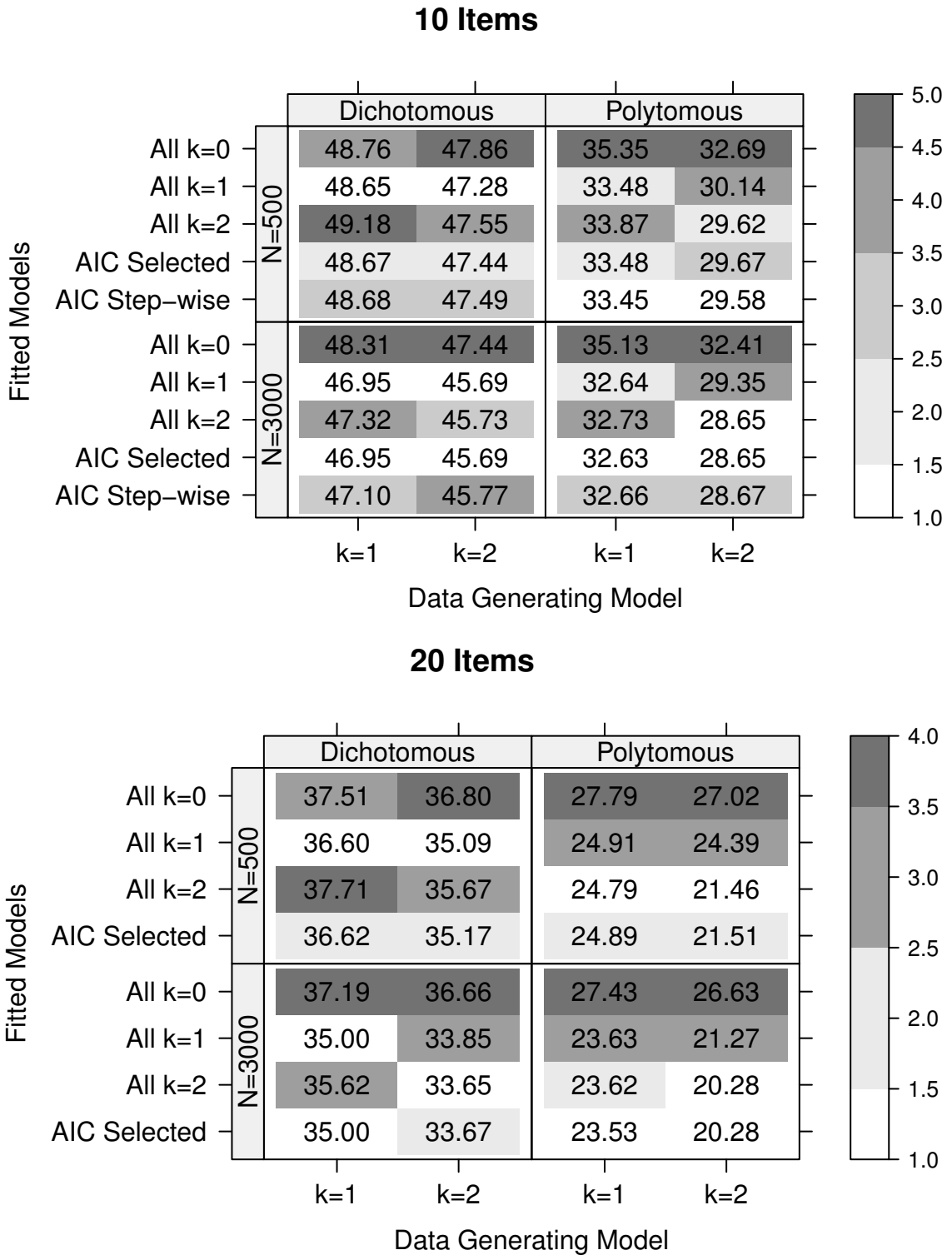
Note. Response curves for higher categories are darker.

Figure 4: RIMSE for Study 1 item response functions



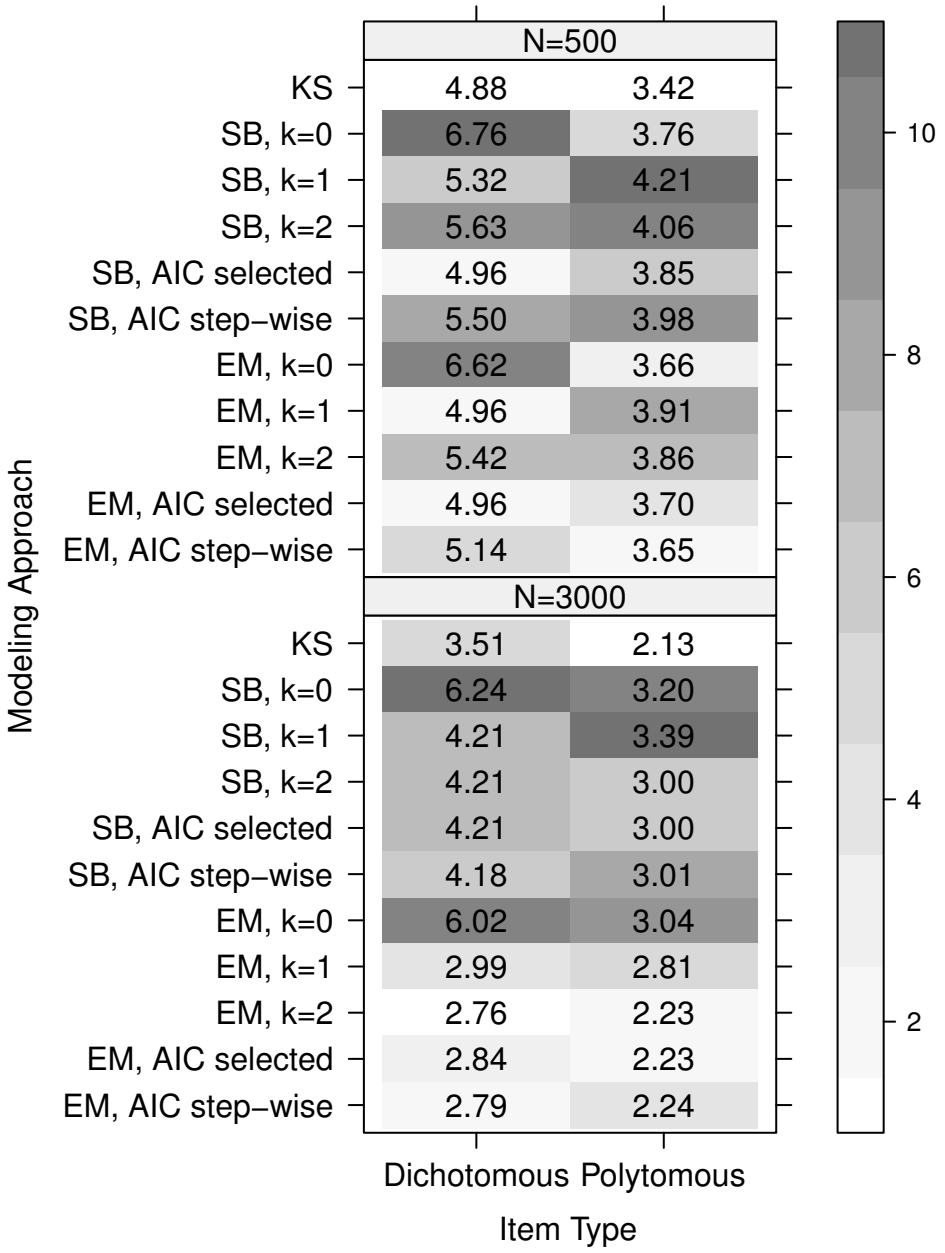
Note. Lower RIMSE values are better. Shading (white = better) indicates rank (1 = best) of each fitted model's performance for a particular data generation combination versus other fitted models.

Figure 5: $RMSE_\theta$ for Study 1 latent trait estimates



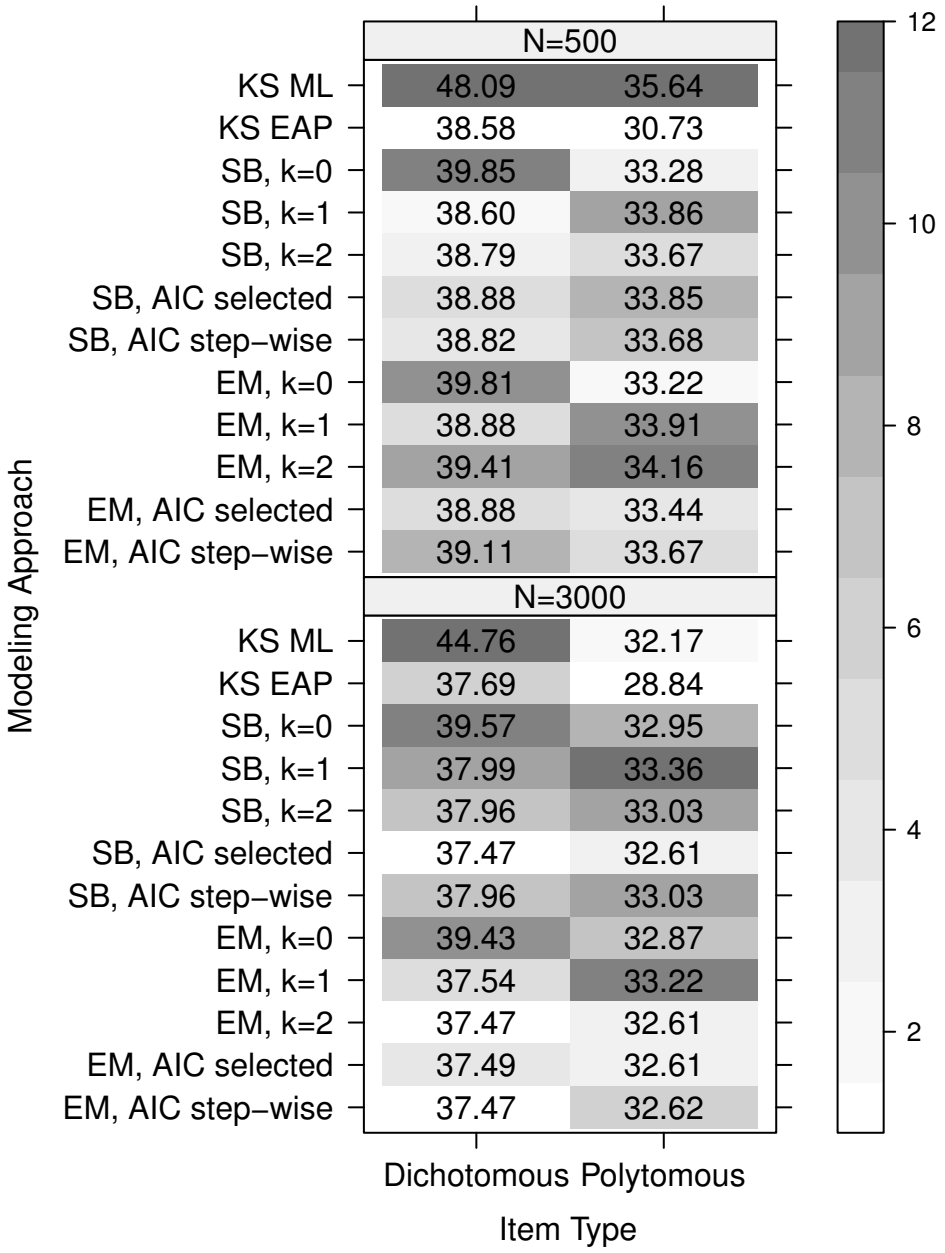
Note. Lower RMSE values are better. Shading (white = better) indicates rank (1 = best) of each fitted model's performance for a particular data generation combination versus other fitted models.

Figure 6: RIMSE for Study 2 item response functions



Note. Lower RIMSE values are better. Shading (white = better) indicates rank (1 = best) of each fitted model's performance for a particular data generation combination versus other fitted models. KS = Kernel smoothing; SB = Surrogate-based estimation; EM = EM MML (Maximum marginal likelihood using the EM algorithm).

Figure 7: RMSE_θ for Study 2 latent trait estimates



Note. Lower RMSE values are better. Shading (white = better) indicates rank (1 = best) of each fitted model's performance for a particular data generation combination versus other fitted models. KS = Kernel smoothing; SB = Surrogate-based estimation; EM = EM MML (Maximum marginal likelihood using the EM algorithm).