The Monotonic Polynomial Graded Response Model: Implementation and a Comparative Study

Carl F. Falk

Abstract

We present a monotonic polynomial graded response (GRMP) model that subsumes the unidimensional graded response model for ordered categorical responses and results in flexible category response functions. We suggest improvements in the parameterization of the polynomial underlying similar models, expand upon an underlying response variable derivation of the model, and in lieu of an overall discrimination parameter we propose an index to aid in interpreting the strength of relationship between the latent variable and underlying item responses. In applications, the GRMP is compared to two approaches: (a) a previously developed monotonic polynomial generalized partial credit (GPCMP) model; and (b) logistic and probit variants of the heteroscedastic graded response (HGR) model that we estimate using maximum marginal likelihood with the expectation–maximization algorithm. Results suggest that the GRMP can fit real data better than the GPCMP and the probit variant of the HGR, but is slightly outperformed by the logistic HGR. Two simulation studies compared the ability of the GRMP and logistic HGR to recover category response functions. While the GRMP showed some ability to recover HGR response functions and those based on kernel smoothing, the HGR was more specific in the types of response functions it could recover. In general, the GRMP and HGR make different assumptions regarding the underlying response variables, and can result in different category response function shapes.

Keywords
graded response model, nonparametric item response theory, monotonic polynomial, heteroscedastic graded response model

A recent flexible approach to response function estimation involves replacing the linear predictor of standard item response models—the two- and three-parameter logistic (2PL and 3PL; Birnbaum, 1968) and generalized partial credit (GPC) models (Muraki, 1992)—with a monotonic polynomial (MP; Falk & Cai, 2016a, 2016b; Liang, 2007; Liang & Browne, 2015). Although there are other flexible models (Duncan & MacEachern, 2013; Miyazaki & Hoshino, 2009; Ramsay & Wiberg, 2017), the MP approach is promising for several reasons. Variants of

1McGill University, Montreal, Quebec, Canada

Corresponding Author:
Carl F. Falk, Department of Psychology, McGill University, 2001 McGill College, 7th Floor, Montreal, Quebec, Canada H3A 1GI.
Email: carl.falk@mcgill.ca
the MP approach that use the expectation–maximization (EM) algorithm to maximize the marginal likelihood (EM-MML; Bock & Aitkin, 1981) can be used with multiple groups, facilitating their use in scaling, linking, or tests of differential item functioning (Falk & Cai, 2016a; Feuerstahler, 2016). Under conditions of high information, the approach has been shown in simulations (Falk & Cai, 2016a, 2016b; Feuerstahler, 2016) to improve recovery of response functions and item fit on par with kernel smoothing (Ramsay, 1991) and smoothed isotonic regression (Y.-S. Lee, 2007). With EM-MML estimation, the MP approach can be used with missing data and can be used on the same test as standard items rather than requiring all items to be estimated nonparametrically.

The main purpose of this manuscript is to define the monotonic polynomial graded response (GRMP) model, formed by replacing the linear predictor of the graded response model (GRM; Samejima, 1969, 1972) with an MP. This development is important, given the popularity of the GRM for psychological assessments. Indeed, some advocate use of a nonparametric model for personality or psychopathology data to reveal how certain items behave differently than what is expected (Meijer & Baneke, 2004).

In addition, previous MP-based models considered logistic building blocks to be monotonically increasing. Here, we propose a modification to the MP parameterization that retains monotonicity, but allows for boundary response functions (BRFs) of the GRMP to be either all increasing or all decreasing for any given item. Such a case is useful when some items are reverse coded and may be useful to multidimensional extensions of the model.

We also seek to enhance interpretation of MP-based models by discussing features of the GRMP as they relate to other variants of graded models such as the heteroscedastic graded response model (HGR; Molenaar et al., 2012), expanding upon a derivation of the model using the underlying variable approach (Bolt, 2005; Takane & de Leeuw, 1987) and providing conversion from logistic to probit parameterizations of the model. Although MP-based models lack a discrimination parameter, this work results in a quantity that may signify how closely related the item is to the latent trait.

We compare the GRMP with the monotonic polynomial generalized partial credit (GPCMP) model (Falk & Cai, 2016a) using data from the Patient Reported Outcomes Measurement Information System (PROMIS; Hansen et al., 2014). Next, the GRMP, GPCMP, and HGR are compared on data from the Synthetic Aperature Personality Assessment (SAPA) project (Condon, 2018). Although the original HGR model considered only a probit link function and directly maximized the marginal likelihood (Bock & Lieberman, 1970) using Mx (Neale et al., 2002), we implement both logistic and probit variants using EM-MML. A test of HGR with EM-MML is important as otherwise the HGR is not feasible for a long test. The two best fitting models from this latter example (logistic HGR and GRMP) are then compared in a small simulation study.

The remainder of this manuscript is organized as follows. In the Proposed Item Model section, we present the GRMP, its alternative parameterizations, and briefly discuss the relationship between the GRMP and other models. In the Empirical Examples section we present both empirical examples. In the Simulations section, we present two small simulation studies. We end in the Discussion and Conclusion with remaining challenges on the use of MP-based item models, and promising avenues for future research.

The Proposed Item Model

GRMP Model

Consider \( i = 1, 2, \ldots, N \) independent respondents who complete a subset of \( j = 1, 2, \ldots, n \) polytomous items with \( k = 0, 1, \ldots, K_j - 1 \) as response options for item \( j \). Let \( Y_{ij} \) be a random
variable for respondent \( i \)'s response to item \( j \), and \( y_{ij} \) its realization, \( y_{ij} \in [0, 1, \ldots, K_j - 1] \).
The latent variable is \( \theta \), which is often assumed standard normal in single group applications.
The GRM (Samejima, 1969, 1972) is a popular choice for representing the relationship between 
\( \theta \) and responses to ordered categorical items. To construct the model, consider first BRFs that
define the probability that a response is equal to or greater than category \( k \):

\[
P_j(Y_{ij} \geq k|\theta_i) = \frac{1}{1 + \exp(-(c_{jk} + a_j\theta_i))} = \Psi(c_{jk} + a_j\theta_i)
\]

where \( \Psi(\cdot) \) is the cumulative distribution function for the standard logistic distribution, and \( a_j \)
is a slope. The intercepts are ordered, \( c_{j1} > c_{j2} > \ldots > c_{j,K_j-1} \), such that BRFs are parallel and do
not cross (top-left of Figure 1). BRFs for the GRM form category response functions (CRFs) by
taking the difference between adjacent BRFs (top-right of Figure 1):

\[
T_j(k|\theta_i) = \begin{cases}
1 - P_j(Y_{ij} \geq (k+1)|\theta_i) & \text{if } k = 0 \\
\frac{P_j(Y_{ij} \geq k|\theta_i) - P_j(Y_{ij} \geq (k+1)|\theta_i)}{P_j(Y_{ij} \geq (k+1)|\theta_i)} & \text{if } 0 < k < K_j - 1 \\
P_j(Y_{ij} \geq (k+1)|\theta_i) & \text{if } k = K_j - 1
\end{cases}
\]

To obtain BRFs for the GRMP, we replace the term \( a_j\theta_i \) in Equation (1) by an MP, \( m(\theta; b_j) \):

\[
P_j(Y_{ij} \geq k|\theta_i) = \frac{1}{1 + \exp(-(c_{jk} + m(\theta_i; b_j)))} = \Psi(c_{jk} + m(\theta_i; b_j))
\]

where \( b_j = [b_{j,1} \ldots b_{j,2q_j+1}] \) is a \( 2q_j + 1 \) vector of coefficients, and \( q_j \) is a user-specified non-negative integer. Suppressing item and respondent subscripts, \( m(\theta; b) \) is of order \( 2q + 1 \), and its derivative (with respect to \( \theta \)), \( m'(\theta; a) \), also contains \( 2q \) coefficients in \( a = [a_{j,1} \ldots a_{j,2q}] \):

\[
m(\theta; b) = b_1\theta + b_2\theta^2 + \ldots + b_{2q+1}\theta^{2q+1}
\]

\[
m'(\theta; a) = a_0 + a_1\theta + a_2\theta^2 + \ldots + a_{2q}\theta^{2q}
\]

Use of Equation (3) results in more flexible BRFs (bottom-left of Figure 1), and Equation (2)
is again used to construct CRFs in a completely analogous manner (bottom-right of Figure 1).

The GRMP is a heterogeneous model (Samejima, 2008, 2010), as BRFs for a single item do
not follow the exact same shape. Bends occur in the same locations along \( \theta \) due to ordered
intercepts and BRFs that are a function of an MP involving \( \theta \). This is in contrast to the GRM,
in which all BRFs differ only in their location along the latent trait. The BRFs for the GRMP
are also asymmetric (S. Lee & Bolt, 2018; Samejima, 2000). The shape of BRFs is not symmetric
about points where its concavity changes, unlike the BRFs in Equation (1).

**MP Parameterization**

The coefficients, \( b_j \), are not directly estimated, but are a function of other parameters. To maintain
monotonicity, the derivative, \( m'(\theta; a) \), has typically been parameterized such that it always
non-negative, and the coefficients of \( m(\theta; b) \) can easily be obtained from this parameterization
(e.g., see Falk & Cai, 2016a; Liang, 2007). In this article, we propose a slight modification to
parameterizations used in the psychometrics literature:

\[
m'(\theta; a) = m'(\theta; \lambda, \alpha, \tau) = \begin{cases}
\lambda \prod_{u=1}^{q} \left(1 - 2\alpha_u\theta + (\alpha_u^2 + \exp(\tau_u))\theta^2\right) & \text{if } q > 0 \\
\frac{\lambda}{\sqrt{\tau}} & \text{if } q = 0
\end{cases}
\]
where \( \alpha = [\alpha_1, \ldots, \alpha_q] \) and \( \tau = [\tau_1, \ldots, \tau_q] \) are vectors of parameters. In a previous study by Falk and Cai (2016a), \( \lambda \) was reparameterized as \( \lambda = \exp(\omega) \). In a study by Liang and Browne (2015), \( \exp(\tau_u) \) was replaced by \( \beta_u \) and \( \lambda \geq 0 \) and \( \beta_u \geq 0 \) for all \( u = 1, \ldots, q \) were required constraints. Here, we place no estimation constraints on \( \lambda \). If using EM-MML for estimation, Falk and Cai’s (2016a) parameterization allows for unconstrained optimization at the M-step, whereas that by Liang and Browne (2015) requires enforcing inequality constraints. The use of either parameterization results in response functions (or logistic building blocks) that are monotonically increasing. This feature is analogous to having all test items with positive loadings, which may make sense for educational tests. However, some noncognitive tests (such as those in personality) regularly employ items that are reverse-keyed. For such tests and before reverse coding, we would expect some response functions to be monotonic decreasing—analogous to having negative loadings. Although we do not yet develop multidimensional models, use of a parameterization that allows the response surface to monotonically increase across the space for one latent trait, but monotonically decrease for another latent trait may be needed. The parameterization in Equation (6) accomplishes this change by allowing \( \lambda \) to be either positive (increasing) or negative (decreasing). Example item parameters for the item in Figure 1 appear in Table 1.

**Alternative Representation and Relationship to Other Approaches**

In a vein analogous to the relationship between categorical factor analysis and item response theory (Takane & de Leeuw, 1987), additional insight into the GRMP can be obtained by considering other related ways of deriving or representing the model. For instance, Falk and Cai
(2016b) considers that underlying observed responses to item $j$, $y_j$, is a continuous variable that is a function of an MP,

$$y_j^* = m(\theta, b_j^*) + \varepsilon_j = \sum_{i=1}^{2q+1} b_{ji} \theta_i + \varepsilon_j$$

(7)

where $m(\theta; b_j^*) = b_{j1} \theta_1 + b_{j2} \theta_2^2 + \cdots + b_{j2q+1} \theta_{2q+1}$ is the polynomial, $b_j^*$ is a 2$q+1$ vector of its coefficients for item $j$ under this alternative parameterization, and $\varepsilon_j$ is an error term.

In this article, we suppose that $y_j^*$ is discretized into $K_j$ categories according to $K_j - 1$ thresholds, $r_j = [r_{j,1} \quad r_{j,2} \cdots \quad r_{j,K_j-1}]$,

$$y_j = \begin{cases} 0 & \text{if } y_j^* < r_{j,1} \\ k & \text{if } r_{j,k} < y_j^* < r_{j,k+1} \\ K_j - 1 & \text{if } y_j^* > r_{j,K_j-1} \end{cases}$$

(8)

Then, let $E[y_j^*|\theta] = m(\theta; b_j^*)$ and $\text{Var}(y_j^*|\theta) = \psi_j^2$ be the conditional expectation and variance of $y_j^*$ given $\theta$, respectively. If we further assume that the errors are normally distributed, $\varepsilon_j \sim \mathcal{N}(0, \psi_j^2)$, then BRFs can be represented in the following way:

$$P_j(Y_{ij} \geq k|\theta_i) = \frac{1}{\psi_j \sqrt{2\pi}} \int_{r_{j,k}}^{\infty} \exp \left( -\frac{1}{2} \left( \frac{y_j^* - m_j(\theta_i; b_j^*)}{\psi_j} \right) ^2 \right) dy_j^* = \Phi \left( \frac{m_j(\theta_i; b_j^*) - r_{j,k}}{\psi_j} \right)$$

(9)

where $\Phi(\cdot)$ is a standard normal cumulative distribution function. CRFs can be constructed from Equation (9) in a manner analogous to that already presented for the GRMP by using Equation (2). Equation (9) resembles the BRF for the GRM with a probit link function (or a standard normal ogive model), except with the linear predictor replaced by an MP.

Falk and Cai (2016b) provide a similar expression to Equation (9), except with a lower asymptote and for dichotomous items. These authors were concerned with priors to prevent $\psi_j$ from becoming too small and potentially causing estimation difficulty. Here, we provide additional details on interpretation of Equation (9) and conversion between logistic and probit parameterizations. For example, it is possible to approximate what the MP coefficients under the GRMP would be if we had assumed normally distributed errors using the fact that $\Psi(z) \approx \Phi(z/D)$, where $D$ is the usual scaling constant (e.g., $D = 1.702$). That is, we can divide all logistic intercepts and MP coefficients by $D$ to convert them to the normal ogive metric, $\tilde{c}_k = c_k/D$, and $\tilde{b} = b/D$, noting that $\Phi(\tilde{c}_k + m(\theta; \tilde{b})) \approx \Phi(c_k + m(\theta; b))$.

Further conversion to the parameterization in Equation (9) is also possible (Table 2). If we constrain the variance of $y_j^*$ to 1, coefficients under the standardized normal ogive metric are interpretable as standardized regression coefficients and can be approximated:

| Parameter | $\lambda$ | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $\alpha_1$ | $\tau_1$ | $\alpha_2$ | $\tau_2$ | $\alpha_3$ | $\tau_3$ |
|-----------|-----------|-------|-------|-------|-------|-------------|---------|-------------|---------|-------------|---------|
| $q=0$     | 1.76      | 4.20  | 2.09  | 0.27  | -1.57 |             |         |             |         |             |         |
| $q=3$     | 1.53      | 4.07  | 1.93  | 0.12  | -1.73 | -0.57       | -9.72   | 0.69        | -3.35   | -0.39       | 0.55    |
Table 2. Example Monotonic Polynomial Coefficients From Different Parameterizations.

|                | Logistic | Normal ogive | Stand. normal ogive |
|----------------|----------|-------------|--------------------|
|                | $q = 0$  | $q = 3$     | $q = 0$            | $q = 3$     | $q = 0$ | $q = 3$ |
| $c_1$          | 4.203    | 4.070       | $r_1$              | 2.469       | 2.391   | $r_1$     | -1.715  | -0.219  |
| $c_2$          | 2.091    | 1.926       | $r_2$              | 1.229       | 1.131   | $r_2$     | -0.854  | -0.104  |
| $c_3$          | 0.273    | 0.122       | $r_3$              | 0.161       | 0.072   | $r_3$     | -0.112  | -0.007  |
| $c_4$          | -1.572   | -1.728      | $r_4$              | -0.924      | -1.015  | $r_4$     | 0.642   | 0.093   |
| $b_1$          | 1.763    | 1.533       | $b_1$              | 1.036       | 0.901   | $b_1$     | 0.719   | 0.083   |
| $b_2$          | 0.400    | 0.235       | $b_2$              | 0.287       | 0.207   | $b_2$     | 0.726   | 0.022   |
| $b_3$          | 0.488    | -0.337      | $b_3$              | -0.198      | -0.162  | $b_3$     | -0.157  | -0.001  |
| $b_4$          | -0.342   | -0.201      | $b_4$              | -0.198      | -0.186  | $b_4$     | -0.186  | -0.018  |
| $b_5$          | -0.337   | -0.198      | $b_5$              | -0.198      | -0.186  | $b_5$     | -0.186  | -0.018  |
| $b_6$          | 0.098    | 0.057       | $b_6$              | 0.039       | 0.004   | $b_6$     | 0.005   | 0.007   |
| $b_7$          | 0.067    | 0.039       | $b_7$              | 0.039       | 0.004   | $b_7$     | 0.005   | 0.007   |

Note. Normal ogive and standardized normal ogive coefficients are approximated from the logistic coefficients using conversion formulae presented in the main text. The coefficients here correspond to the item that also appear in Table 1 and Figure 1.

\[
\begin{align*}
    r_{j,k} &= \frac{-c_{j,k}/D}{\sqrt{1 + (1/D^2)b_j^{\top}\Gamma b_j}}; \\
    b_{j,t} &= \frac{b_{j,t}/D}{\sqrt{1 + (1/D^2)b_j^{\top}\Gamma b_j}} \\
    \hat{r}_{j,k} &= \frac{-\tilde{c}_{j,k}}{\sqrt{1 + b_j^{\top}\Gamma b_j}}; \\
    \tilde{b}_{j,t} &= \frac{\tilde{b}_{j,t}}{\sqrt{1 + b_j^{\top}\Gamma b_j}} \\
\end{align*}
\]

where $b_j^{\top}\Gamma b_j = \text{var}(m(\theta; b_j))$, and in which $\Gamma$ is a symmetric matrix containing the variance–covariances among the $2q + 1$ latent variable terms in $m(\theta; b_j)$ (see Supplementary Materials). Had the model been estimated using the unstandardized or standardized normal ogive variant, conversion between these two parameterizations is also possible:

\[
\begin{align*}
    r_{j,k} &= \frac{-\hat{c}_{j,k}}{\sqrt{1 + b_j^{\top}\Gamma b_j}}; \\
    \hat{b}_{j,t} &= \frac{\hat{b}_{j,t}}{\sqrt{1 + b_j^{\top}\Gamma b_j}} \\
    \tilde{c}_{j,k} &= \frac{-\hat{r}_{j,k}}{\sqrt{1 - b_j^{\top}\Gamma b_j}}; \\
    \tilde{b}_{j,t} &= \frac{\hat{b}_{j,t}}{\sqrt{1 - b_j^{\top}\Gamma b_j}} \\
\end{align*}
\]

Equation (7) resembles a polynomial regression, with a zero intercept and the additional constraint that the polynomial is monotonic. Description of the model may also rely on this analogy: Had none of the quantities in Equation (7) been latent and the polynomial had no special constraints, we could use standard software for linear regression to obtain the coefficients. The GRMP is modeling a nonlinear but monotonic relationship between the latent variable, $\theta$, and the underlying response variable, $y_j^\ast$. The thick solid line in the left panel of Figure 2 displays this nonlinear relationship between $\theta$ and $y_j^\ast$, using the $q = 3$ polynomial in Table 2 on the standardized normal ogive metric. This representation may have useful substantive interpretations to the extent that we would expect this relationship to be stronger at some regions of the latent trait than at others.

Differences between the GRMP and other models are also easier to see. For example, Molenaar et al. (2012) present BRFs for the HGR equivalent to

\[
P_j(Y_{ij} \geq k|\theta_i) = \Phi\left(\frac{y_j^\ast + b_j^\ast\theta_i - r_{j,k}}{\psi_j^{\prime}\theta_i}\right)
\]
where \( q_j^2 = 2\delta_j/(1 + \exp(-\delta_j(\theta_i - E(\theta))/SD(\theta))) \) is a function of \( \theta \). When \( \delta_j \) is nonzero, the HGR allows for heteroscedastic errors (right panel of Figure 2). Such a model also results in asymmetric BRFs and CRFs with nonstandard shapes (e.g., Molenaar et al., 2012, Figure 7, p. 475). In contrast, the asymmetric BRFs under the GRMP are assumed to be due to a nonlinear relationship between the latent trait, \( \theta \), and the underlying response variable, \( y_j^* \). Errors are assumed homoscedastic and either normal or logistic, depending on the link function.

We assume that \( \theta \) is normal, but note that Equation (9) is conditional on \( \theta \), and the response variable itself, \( y_j^* \), need not be normal when \( q > 0 \). In contrast to the GRMP, linearity between \( \theta \) and \( y_j^* \) is assumed under the HGR, which may in some cases be too simple. In the context of MP models, Feuerstahler (2016, 2019) discussed how an equivalent model may be specified in which a monotonic transformation of the latent trait results in changes to MP coefficients for the items. Thus, it is possible to transform away some of the nonlinearity between \( \theta \) and \( y_j^* \) by considering a non-normal distribution for \( \theta \). A transformation for \( \theta \) may imply similar changes across items and may not fully explain all nonstandard CRFs to the extent that there is heterogeneity across items. The main application of this previous work is the recognition of “parameters” for MP approaches, which are not easily interpretable, but allow for linking as well as transformation of the latent trait, CRFs, and information functions to an alternative metric of choice. For example, even if a normal distribution was used for calibration, MP-based models could, in theory, be transformed onto a metric that has the same range as sum scores (the metric used by Ramsay & Wiberg, 2017), which may be palatable by stakeholders and result in changes in information toward the endpoints of the continuum. This feature of MP-based items is both interesting, but a limitation due to indeterminacy of the underlying latent metrics and response functions. One must rely on substantive theory and/or the test purpose to make an appropriate modeling choice, and such decisions are nontrivial.

Finally, in the standard normal ogive model, there is an item parameter (\( b_{j,1}^* \) when \( q_j = 0 \)) that can be interpreted as the correlation between \( \theta \) and \( y_j^* \), and when squared the proportion of variance in \( y_j^* \) that is explained by \( \theta \) (Lord, 1980). MP-based item models lack such a parameter and also lack a single-item discrimination parameter. However, some useful information about the strength of relationship between \( \theta \) and individual responses can be gleaned from

![Figure 2. Relationship between latent trait and underlying response for GRMP and HGR, with density for three conditional distributions.](image)

Note. GRMP = monotonic polynomial graded response; HGR = heteroscedastic graded response.
approximations to $\psi_j^2$. If $y_j^*$ is assumed to have unit variance, then $\psi_j^2$ is interpretable as an estimate of the unexplained variance in $y_j^*$. Conversely, $\rho_j^2 = 1 - \psi_j^2$ corresponds to proportion of variance in the underlying response variable that is explained by $\theta$, otherwise known as the coefficient of determination. Thus, estimation of $\psi_j^2$ or $\rho_j^2$ may be useful for substantive interpretation. The denominator in Equation (10) is an approximation due to

$$\sigma^2 = \frac{1}{1 + (1/D^2) b_j^T \Gamma b_j}.$$  

If estimating the normal ogive variants directly, then $\psi_j^2$ is available in terms of parameters from both versions, $\psi_j^2 = 1/(1 + b_j^T \Gamma b_j) = 1 - b_j^T \Gamma b_j$. Under the $q = 0$ and $q = 3$ examples already given, $\psi^2 \approx 0.482$ and $\psi^2 \approx 0.008$, respectively, corresponding to $\rho_j^2 \approx 0.518$ and $\rho_j^2 \approx 0.992$ variance explained in $y_j^*$.

**Empirical Examples**

**Example 1: PROMIS**

The purpose of the first application was to compare the GRMP and GPCMP. We were particularly interested in (a) whether use of an MP (greater than $q = 0$) for the GRMP tended to improve model fit; (b) whether the GRMP tended to fit better/worse than the GPCMP; and (c) whether the GRMP resulted in similar CRF shapes as the GPCMP.

**Data.** We used responses to sixteen 5-category Likert-type items from 3,605 daily smokers for measuring hedonic benefits of tobacco smoking from PROMIS (Hansen et al., 2014). These data overlap with that analyzed by Falk and Cai (2016a), who found better fit (according to Akaike information criterion [AIC]) with the GPCMP using at least $q = 1$ polynomials for all items. We examined whether a similar result would occur for the GRMP. The study employed 27 overlapping test forms, and approximately 35% of the data we examine here were missing.

**Estimation and fitted models.** Ten final models were estimated: five for the GRMP and five for the GPCMP. We first describe the GRMP. Three models were estimated in which the MP order was the same for all items ($q = 0$, $q = 1$, and $q = 2$). For these models, estimation of the lowest order polynomial, $q = 0$, was done first, with a small positive value, $\exp(-.5) \approx .61$, as the starting value for $\lambda$ and equally spaced values between 1.5 and $-1.5$ for intercept starting values. Estimates for $\lambda$ and intercepts were used as starting values for corresponding parameters when $q = 1$, with $\alpha_j$ started at 0, and $\tau_1$ at $-2.3$. The same strategy was used when moving from $q = 1$ to $q = 2$: Estimates of $\lambda$, intercepts, $\tau_1$, and $\alpha_1$ from the $q = 1$ model were used as starting values (with $\tau_2$ and $\alpha_2$ started at zero).

An additional two models employed a stepwise approach using either AIC or Bayesian information criterion (BIC), which we refer to as AIC-step and BIC-step, respectively. In particular, these models started with all items at the lowest order polynomial ($q = 0$). We then fit $n$ models that separately considered increasing $q$ by 1 for each item, and selected the best improved model according to AIC or BIC. This process repeated until no improvement could be found, up to a maximum of $q = 3$ for each item ($\tau_3$ and $\alpha_3$ were started at zero).

The GPCMP model was based on Falk and Cai (2016a). We used starting values of $\omega$, and starting values for intercepts ordered in the opposite direction from $-1.5$ to 1.5. All other estimation details were identical to that already presented for the GRMP.

We used Bayes modal estimation coupled with the EM algorithm (Mislevy, 1986), with integrals evaluated using rectangular quadrature from $-5$ to 5 in .1 increments. A standard normal $\theta$ was assumed for all models fit in this manuscript. Following Falk and Cai (2016a), we used weak priors, $\pi(\alpha_u) \sim N(0, 500)$ and $\pi(\tau_u) \sim N(-1, 500)$ to stabilize estimation. Evaluation of AIC and BIC was done by plugging in obtained estimates into the marginal log-likelihood (Mislevy, 1986). A maximum of 500 and 2,000 iterations and relative tolerance of $1.0 \times 10^{-9}$
and $1.0 \times 10^{-7}$ was set for the M-step and E-step, respectively. At the M-step, a Newton–Raphson algorithm with analytical derivatives was used for estimation (see Supplemental Materials). We used a modified version of the *rpf* package (Pritikin, 2016) for CRFs and derivatives, and estimated models using *OpenMx* (Neale et al., 2016; Pritikin et al., 2015).

### Results

Whether MP-based models improved fit depended slightly on the information criterion (Table 3). The AIC-stepwise model always fit best according to AIC, followed by either the $q = 1$ (GRMP) or $q = 2$ (GPCMP) models. Examination of $q$ under the AIC-stepwise models (see Supplementary Materials) also suggested that improvement in fit was possible with all items as $q = 1$ or higher (some up to $q = 3$). In contrast, BIC favored models with few increases in polynomial order with the BIC-stepwise model favored for the GPCMP, followed by all items with $q = 1$. For the GRMP, the BIC-stepwise approach did not find a model that improved fit beyond the GRM. With the exception of this model, all BIC-stepwise models had items modeled as $q = 0$ or $q = 1$. The GRMP also tended to fit better than the GPCMP; even the worst fitting of any GRMP model fit better than the best fitting GPCMP model. Stepwise models involving the GPCMP tended to suggest more higher order polynomials than did stepwise models with the GRMP. Despite these differences between the GRMP and GPCMP, the two-item models tended to result in CRF shapes that were very similar (Figure 3).

Finally, we compared estimates of $\rho_j^2$ from the simplest of estimated models (the all $q = 0$ model) versus the most complex model preferred by AIC (the AIC-stepwise model). There was a tendency for $\rho_j^2$ to increase when higher order polynomials were used. In the most extreme case, this relationship jumped from .41 to .99 for item 16. When polynomial order did not increase, a change in $\rho_j^2$ was not apparent (see Supplementary Materials).

### Example 2: SAPA

In Example 2, we continue to compare the GRMP and GPCMP and add the HGR model. Thus, we were interested in whether the probit or logistic variants of the HGR would yield better fit than the GRMP, and whether similar CRF shapes would be observed.

### Data

We examined 16,127 responses to ten 6-category anxiety items from the HEXACO Personality Inventory-Revised (HEXACO-PI-R; K. Lee & Ashton, 2018). The data were collected as part of the SAPA project (Condon, 2018; Condon & Revelle, 2015). We chose anxiety rather than the entire emotionality dimension to ensure that MP-based models are less likely to pick up on any local dependence that may arise from a multidimensional measurement.

### Table 3. Obtained AIC and BIC for PROMIS Data.

| Dataset/item model | $q = 0$ | $q = 1$ | $q = 2$ | AIC-step | BIC-step |
|--------------------|---------|---------|---------|-----------|-----------|
| AIC                |         |         |         |           |           |
| GRMP               | 96,042  | 95,963  | 95,975  | 95,939    | 96,042    |
| GPCMP              | 96,667  | 96,455  | 96,440  | 96,406    | 96,522    |
| BIC                |         |         |         |           |           |
| GRMP               | 96,537  | 96,656  | 96,866  | 96,744    | 96,537    |
| GPCMP              | 97,162  | 97,148  | 97,332  | 97,248    | 97,129    |

Note. Two best fitting models for each row are in bold. AIC = Akaike information criterion; BIC = Bayesian information criterion; AIC-step = AIC-stepwise model; BIC-step = BIC-stepwise model; PROMIS = Patient Reported Outcomes Measurement Information System; GRMP = monotonic polynomial graded response; GPCMP = monotonic polynomial generalized partial credit.
instrument. The SAPA project employs a planned missing data design and responses to items are sparse within this large sample: only 2,666 participants on average completed each of the items we examined here (83% missing data). Finally, the anxiety scale included five reverse-keyed items, allowing us to test the GRMP with negatively loading items. In what follows, such items were reverse coded only when comparing the older MP parameterization of the GPCMP against the GRMP.

**Fitted models.** The same GRMP and GPCMP models were fit as in the first example. For the HGR (Molenaar et al., 2012), a probit or logistic link function can be used in Equation (13), which we refer to as HGR-P and HGR-L, respectively, with optionally scaling the entire contents of the HGR-L by 1.702 to ensure that item parameters are approximately on the same metric as the HGR-P. We programmed CRFs for the HGR-P and HGR-L and used the createItem() function of the mirt package (Chalmers, 2012) for EM-MML estimation with numerical derivatives.\(^4\) Starting values following supplementary code from Molenaar et al. (2012) were used, with \(r_{j,0} = -2\) and \(r_{j,1} = -1\) fixed for identification. For notational similarity to the GRMP, we use \(h_j = 1\) to indicate modeled heteroscedasticity for any given item, and \(h_j = 0\) as homoscedastic. Versions of the HGR-P and HGR-L were estimated in which all items were modeled with heteroscedastic errors (i.e., all \(h = 1\)). We also estimated AIC and BIC-stepwise versions of both models in which heteroscedastic errors were added to each item one at a time (\(h\) may vary across items).

Although the HGR-P could be checked against Mx (Neale et al., 2002) with code by Molenaar et al. (2012), we encountered some discrepancies across programs and some estimation difficulty with Mx. Since Mx was slow to estimate relative to our implementation, we instead decided to conduct the simulation study that follows this empirical example as an

---

**Figure 3.** Category response functions for PROMIS data, GRMP (thick, solid, and light) and GPCMP (thin, dashed, and dark) AIC-stepwise models.

Note. Darker lines indicate a higher category. Inside each panel, \(q\) for GRMP appears on the left and \(q\) for GPCMP on the right. PROMIS = Patient Reported Outcomes Measurement Information System; GRMP = monotonic polynomial graded response; GPCMP = monotonic polynomial generalized partial credit; AIC = Akaike information criterion.
Results. Results for the GRMP and GPCMP were similar to that in the first empirical example: The GRMP tended to fit better, some items were modeled with at least $q = 1$ for all stepwise models, and stepwise approaches tended to suggest fewer higher order polynomials for the GRMP than the GPCMP (Table 4). Results with the HGR models were mixed. Stepwise versions of the GRMP fit better than the HGR-P according to both AIC and BIC. However, the stepwise versions of the HGR-L exhibited the best fit of any fitted models. Finally, modeling all items with heteroscedastic errors was not necessarily preferred by BIC but yielded similar fit to stepwise approaches according to AIC.

The GRMP correctly estimated CRFs for the reverse-keyed items (1–2, and 7–9) as $\lambda$ estimates were negative, and the highest category CRF increased in probability at lower levels of $u$. Keying all items in same direction, we found some discrepancies in CRF shapes across the GRMP and HGR-L. Figure 4 plots CRFs for these approaches from both AIC-stepwise models. Items 3, 5, and 7 were modeled with $q = 0$ and $h = 0$, and both models had similar shapes. However, Item 6 was modeled with $q = 0$ by the GRMP, but HGR-L suggested a large amount of heteroscedasticity. Items 9 and 10 also had nonstandard CRFs, but the shape appeared to be different depending on the item model. Integrating over $X = 121$ quadrature nodes from $-6$ to 6, we calculated root integrated mean square differences, \( \text{RIMSD} = \frac{1}{n} \sum_{j=1}^{n} \left( \frac{\text{ES}_j(\theta)}{\text{ES}_j(\theta)} \right)^2 \), as a measure of the root sum of squared discrepancy between expected scores, \( \text{ES}_j(\theta) = \frac{1}{n} \sum_{j=1}^{n} \phi(\theta_j) \). For the GRMP (\( \text{ES} \)) and HGR-L (\( \tilde{\text{ES}} \)). RIMSD had a similar pattern to these visual inspections, as the values for the 10 items are as follows (larger values indicate a greater difference): 0.036, 0.031, 0.002, 0.026, 0.003, 0.037, 0.001, 0.021, 0.060, and 0.045.

Simulations

Simulation 1

To check recovery of the GRMP and HGR-L models, and also probe whether either approach can recover CRFs from the opposing model, a simulation study was conducted.
Method. Data were generated for 10 items using item parameters from one of two models: The GRMP or HGR-L AIC-stepwise models from the second empirical example (see Supplementary Materials for item parameters). A standard normal latent trait was assumed with 50 replications and $N=15,000$ per replication. Two versions of each data set were also constructed—one with complete data, and another in which 80% missing completely at random data was induced such that each respondent had completed only two items. We therefore mimicked the second empirical example but also study performance of each approach under high information (i.e., complete data).

To each data set, five models were fit to the data: (1) GRM as natively available in mirt (GRM); (2) GRM custom programmed using the parameterization in Equation (13) and estimated with mirt (GRMcust; $\delta_{i,1}=0$); (3) the GRMP AIC-step model (GRMP); (4) the HGR-L AIC-step model (HGR-L); and (5) HGR-L with all heteroscedastic items (HGR-L all). Examination of Models 1 and 2 allows for a sanity check on our implementation of the HGR-L model as Model 2 should essentially match Model 1. Models 3 and 4 correspond to the main comparisons of interest and may outperform Models 1 and 2. Finally, Model 5 studies the consequences of potentially overfitting using the HGR-L model.

Results. Recovery of CRFs was calculated for each item and replication using root integrated mean square error (RIMSE; e.g., Falk & Cai, 2016a), which is calculated in the same manner as
RIMSD, but computed using expected scores based on the true (ES) and estimated (cES) CRFs. Lower values indicate better CRF recovery. In what follows, RIMSE was additionally averaged over items and replications in each cell of the design (Table 5).

Under missing data, GRMP and HGR-L with AIC-stepwise selection improved recovery over the GRM, but only if the model corresponded to the data generating mechanism. For example, fitting the GRMP to HGR-L data did not tend to improve recovery and vice versa. Under complete data, a similar pattern held with one exception. Fitting the closest model to the true data generating model tended to have the best CRF recovery. However, the GRMP showed some gains in CRF recovery (RIMSE = .008) over the GRM (RIMSE = .013) when fit to HGR-L data, but was not as good as the HGR-L (RIMSE = .003). The opposite was not true in that the HGR-L did not have much of an advantage over the GRM in recovering CRFs from GRMP data.

**Simulation 2**

The second study had two goals. First, based on reviewer comments, we sought to examine CRF recovery from a data generating model that was neither derived from the GRMP nor logistic HGR. Second, MP models typically do not perform well unless very large sample sizes are used. Here, we experimented with a relatively smaller sample size, but with stronger priors for the GRMP in an attempt to better stabilize estimation.

**Method.** To construct a data generating model, a nonparametric approach that also works with incomplete item responses as in our empirical examples is not immediately apparent. We therefore used all respondents with complete data (N = 1,068) from the PROMIS example and estimated CRFs using default options with the KernSmoothIRT package (Mazza et al., 2013) and a polytomous item Kernel smoothing approach (Santor et al., 1994; see Supplementary Materials for CRFs). For each sample size condition (N = 250 and N = 2,500), we generated 50 complete data sets based on these CRFs and using a standard normal θ.

All models described in Simulation 1 were fit to each data set, with one change to the GRMP: Either weak, \( \pi(\alpha_u) \sim N(0, 500) \) and \( \pi(\tau_u) \sim N(-35, 500) \), or strong priors, \( \pi(\alpha_u) \sim N(0, .005) \) and \( \pi(\tau_u) \sim N(-35, .005) \), were used. A value of 0 for \( \alpha_u \) and a large negative value of \( \tau_u \) would result in a GRMP model that reduces to the GRM. The strong priors are aimed at reducing overfitting, and the prior mean implies that our best prior knowledge is that the model is a GRM and not something more complex.

**Results.** The GRMP and strong priors resulted in the best CRF recovery at the smaller sample (RIMSE = .0343; Table 6), though gains above the GRM were very small (RIMSE = .0351). The

| Table 5. RIMSE for Simulation Study I. |
|---------------------------------------|---------------------------------|---------------------------------|-------------------------------|-------------------------------|
| **Fitted model**                      | **Data generating model**       | **GRM**                         | **GRMcust**                   | **GRMP**                      | **HGR-L**                     | **HGR-L all**                 |
|---------------------------------------|---------------------------------|---------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| GRMP (missing)                        | 0.0237                          | 0.0237                          | 0.0156                        | 0.0252                        | 0.0261                        |
| GRMP (complete)                       | 0.0229                          | 0.0229                          | 0.0040                        | 0.0227                        | 0.0227                        |
| HGR-L (missing)                       | 0.0161                          | 0.0161                          | 0.0170                        | 0.0104                        | 0.0103                        |
| HGR-L (complete)                      | 0.0129                          | 0.0129                          | 0.0083                        | 0.0030                        | 0.0031                        |

Note. RIMSE = root integrated mean square error; GRM = graded response model; GRMP = monotonic polynomial graded response; HGR-L = heteroscedastic graded response model with logistic link function.
GRMP was also the best method at the larger sample size, but the difference between different priors essentially disappeared. Both versions of the HGR, in contrast, had worse recovery than the GRM and GRMP at both sample sizes.

Discussion and Conclusion

We have presented the GRMP, an extension of the GRM in which the linear predictor is replaced with an MP. The GRMP can model a nonlinear monotonic relationship between the latent variable and underlying response variables, and can fit data better than the GPCMP, a previously developed MP-based item model (Falk & Cai, 2016a). That the GRMP fits better than the GPCMP is consistent with Maydeu-Olivares (2005), who found that the GRM tended to fit data better than the GPC across several personality data sets. Still, Maydeu-Olivares (2005) suggested that simpler parametric models may be more likely to cross-validate. One reason that some complex item models may improve fit over parametric models stems from possible unmodeled multidimensionality. This is part of the impetus for changes in the MP parameterization, which takes a step toward the development of multidimensional MP models. However, other parameterizations of MPs may also be worthwhile to investigate (e.g., Murray et al., 2013).

In another example, the GRMP fit better than the HGR, but only if using a probit link function as originally presented by Molenaar et al. (2012). “Fit” here is quantified using AIC and BIC, which have their own limitations. For example, BIC did not always suggest that higher order polynomials improved fit, yet BIC has a tendency toward parsimony. In some cases, BIC prefers item response models that are simpler than the true model (Waller & Feuerstahler, 2017). Falk and Cai (2016b) found a similar result when using sum score–based item fit statistics (Orlando & Thissen, 2000) for determining polynomial order: BIC suggested that the use of MPs might not improve fit. Results of simulations, however, suggest that AIC selection and the procedure of Falk and Cai (2016b) or the use of other search algorithms (Falk, 2019) can improve recovery of response functions when using MP-based item models.

Simulations suggested that the GRMP is flexible and can recover a variety of CRF shapes. In the second simulation study, recovery was improved at small sample sizes (N = 250). This result may be due to strong Bayesian priors and remains to be replicated. If such a result holds more generally, it is potentially important, as MP-based models have not been shown to perform well in all but very large samples. It may be that the HGR in this second simulation was overfitting or encountered estimation difficulty and that placing a strong prior on δ1 (centered around zero) is warranted. In addition, as seen in the results of empirical examples, p_j^2 for some items approached one. Although it is not clear if this is problematic, it is possible to adapt priors developed by Falk and Cai (2016b) for preventing ps_j^2 from reaching zero.

That GRMP and HGR tended to suggest different CRF shapes highlights the need for better theory in choosing an appropriate modeling strategy as mere “curve-fitting” (Samejima, 2008)

| Sample size | Fitted model | GRM | GRMcust | GRMP (strong) | GRMP (weak) | HGR-L | HGR-L all |
|-------------|--------------|-----|---------|--------------|-------------|-------|----------|
| 2,250       |              | 0.0351 | 0.0351 | 0.0343       | 0.0351      | 0.0384 | 0.0389   |
| 2,500       |              | 0.0259 | 0.0259 | 0.0221       | 0.0220      | 0.0289 | 0.0289   |

Note. RIMSE = root integrated mean square error; GRM = graded response model; GRMP = monotonic polynomial graded response; HGR-L = heteroscedastic graded response model with logistic link function.

Table 6. RIMSE for Simulation Study 2.
may not help us arrive at a useful model. Strong theory coupled with improved fit may provide a good argument for one model and can result in better recovery of CRFs. This, in turn, we would expect to result in better recovery of latent traits (e.g., Falk & Cai, 2016a). There are a number of reasons why heteroscedasticity in the response variable may occur in practice (Molenaar et al., 2012). The GRMP is apparently more flexible than the HGR and may help uncover other oddities in response functions but at the cost of more estimated parameters. The HGR has one extra parameter per item, whereas the GRMP will have two or more depending on the order of the polynomial. This might suggest that the GRMP is a more complex and less parsimonious model than the HGR. We have therefore chosen to contrast the underlying assumptions regarding the GRMP and HGR. Regardless, this work takes steps at making the HGR more feasible as simulations supported EM-MML estimation for both the GRMP and logistic HGR. Evidence that overall model fit is improved due to either item models may warrant further inspection of item content and an investigation of the possible cause.

Acknowledgment
The author thanks the editors and reviewers for helpful comments during the review process.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study received support from the Fonds de recherche du Quebec—Nature et technologies (grant no. 2019-NC-255344).

ORCID iD
Carl F. Falk https://orcid.org/0000-0002-4788-7206

Supplemental Material
Supplementary material is available for this article online.

Notes
1. Similar τ starting values are used by Murray et al. (2013) and Turlach and Murray (2019).
2. This includes the HGR in subsequent sections.
3. Available at https://github.com/falkcarl/rpf
4. Preliminary attempts at estimation with a skewed latent trait yielded negligible change in model fit.
5. As the nonparametric CRFs are estimated along a grid and there are no parameters, B splines were used to interpolate when determining conditional CRFs at different points along $\theta$ for the purposes of generating item responses.

References
Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 395–479). 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.

Bolt, D. M. (2005). Limited and full information estimation of item response theory models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 27–71). Lawrence Erlbaum.

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

Condon, D. M. (2018). *The SAPA personality inventory: An empirically-derived, hierarchically-organized self-report personality assessment model*. https://doi.org/10.31234/osf.io/sc4p9

Condon, D. M., & Revelle, W. (2015). *Selected personality data from the SAPA-project: 08Dec2013 to 26Jul2014* [Data set]. Harvard Dataverse. http://doi.org/10.7910/DVN/SD7SVE

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). Routledge.

Falk, C. F. (2019). Model selection for monotonic polynomial item response models. In M. Wiberg, S. Culpepper, R. Janssen, J. Gonzalez, & D. Molenaar (Eds.), *Quantitative psychology: The 83rd annual meeting of the psychometric society* (pp. 75–85). Springer Nature.

Falk, C. F., & Cai, L. (2016a). Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis. *Psychometrika, 81*, 434–460.

Falk, C. F., & Cai, L. (2016b). Semi-parametric item response functions in the context of guessing. *Journal of Educational Measurement, 53*, 229–247.

Feuerstahler, L. M. (2016). *Exploring alternate latent trait metrics with the filtered monotonic polynomial IRT model* [Doctoral thesis]. Department of Psychology, University of Minnesota.

Feuerstahler, L. M. (2019). Metric transformations and the filtered monotonic polynomial item response model. *Psychometrika, 84*, 105–123.

Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. G., & Edelen, M. O. (2014). Methodology for developing and evaluating the PROMIS® smoking item banks. *Nicotine & Tobacco Research, 16*(Suppl. 3), S175–S189.

Lee, K., & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment, 25*, 543–556.

Lee, S., & Bolt, D. M. (2018). Asymmetric item characteristic curves and item complexity: Insights from simulation and real data analyses. *Psychometrika, 83*, 453–475.

Lee, Y.-S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement, 31*, 121–134.

Liang, L. (2007). *A semi-parametric approach to estimating item response functions* [Doctoral thesis]. 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. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.

Maydeu-Olivares, A. (2005). Further empirical results on parametric versus non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research, 40*, 261–279.

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

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9*(3), 354–368.

Mishlevy, R. J. (1986). Bayes modal 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.

Molenaar, D., Dolan, C. V., & de Boeck, P. (2012). The heteroscedastic graded response model with a skewed latent trait: Testing statistical and substantive hypotheses related to skewed item category functions. *Psychometrika, 77*(3), 455–478.
Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176.

Murray, K., Müller, S., & Turlach, B. A. (2013). Revisiting fitting monotone polynomials to data. *Computational Statistics, 28*, 1989–2005.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2002). *Mx: Statistical modeling* (6th ed.). Virginia Commonwealth University.

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kickpatrick, R. M., & Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika, 81*, 535–549.

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

Pritikin, J. N. (2016). *rpf: Response probability functions*. https://CRAN.R-project.org/package=rpf

Pritikin, J. N., Hunter, M. D., & Boker, S. M. (2015). Modular open-source software for item factor analysis. *Educational and Psychological Measurement, 75*(3), 458–475.

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

Ramsay, J. O., & Wiberg, M. (2017). A strategy for replacing sum scoring. *Journal of Educational and Behavioral Statistics, 42*(3), 282–307.

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

Samejima, F. (1972). A general model of free-response data. *Psychometric Monographs, 18*.

Samejima, F. (2000). Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves. *Psychometrika, 65*, 319–335.

Samejima, F. (2008). Graded response model based on the logistic positive exponent family of models for dichotomous response. *Psychometrika, 73*, 561–578.

Samejima, F. (2010). The general graded response model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Developments and applications* (pp. 77–107). Taylor & Francis.

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.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393–408.

Turlach, B. A., & Murray, K. (2019). *MonoPoly: Functions to fit monotone polynomials*. https://rdrr.io/cran/MonoPoly/

Waller, N., & Feuerstahler, L. M. (2017). Bayesian model estimation of the four-parameter item response model in real, realistic, and idealized data sets. *Multivariate Behavioral Research, 52*, 350–370.