Educational and Psychological
Measurement
2022, Vol. 82(1) 57–75

DOI: 10.1177/00131644211014261
journals.sagepub.com/home/epm
SAGE

On the Performance of Semi- and Nonparametric Item Response Functions in Computer Adaptive Tests

Carl F. Falk[1] **and Leah M. Feuerstahler**[2]

Abstract

Large-scale assessments often use a computer adaptive test (CAT) for selection of items and for scoring respondents. Such tests often assume a parametric form for the relationship between item responses and the underlying construct. Although semi- and nonparametric response functions could be used, there is scant research on their performance in a CAT. In this work, we compare parametric response functions versus those estimated using kernel smoothing and a logistic function of a monotonic polynomial. Monotonic polynomial items can be used with traditional CAT item selection algorithms that use analytical derivatives. We compared these approaches in CAT simulations with a variety of item selection algorithms. Our simulations also varied the features of the calibration and item pool: sample size, the presence of missing data, and the percentage of nonstandard items. In general, the results support the use of semi- and nonparametric item response functions in a CAT.

Keywords

computer adaptive test, nonparametric IRT, monotonic polynomial, large-scale testing

Introduction

Despite Stout (2001) declaring that nonparametric item response theory (IRT) is viable for the scaling of educational and psychological tests, significant barriers

[1] McGill University, Montreal, Quebec, Canada
[2] Fordham University, Bronx, NY, USA

Corresponding Author:
Carl F. Falk, Department of Psychology, McGill University, 2001 McGill College, 7th Floor, Montreal, QC H3A 1G1, Canada.
Email: carl.falk@mcgill.ca

remain to the use of these approaches. For example, large-scale assessments are more likely to use models such as the two-parameter logistic (2PL), three-parameter logistic (3PL; Birnbaum, 1968), or generalized partial credit model (Muraki, 1992) when modeling the relationship between a latent trait and item responses. These parametric models are often used despite the fact that nonparametric and flexible IRT approaches typically make fewer restrictive assumptions. One challenge to the use of more flexible modeling approaches is the desire to use estimated item response functions (IRFs) in a computer adaptive test (CAT).

The performance of flexibly estimated response functions when used in a CAT is largely unknown as many techniques are not easily tractable with existing testing programs. Classical CAT item selection algorithms, such as maximum Fisher information (MFI; Lord, 1980) or maximum weighted posterior information (MPWI; van der Linden, 1998) are still heavily used in operational CAT settings. These techniques typically require derivatives of the IRF with respect to the latent trait, which are not always readily available for flexibly estimated IRFs. Thus previous applications of CATs to nonparametric IRFs have used alternative item selection algorithms, as otherwise numerical derivatives may be unstable (Y.-P. Chang et al., 2019; Xu & Douglas, 2006).

A novel monotonic polynomial (MP) approach to IRF estimation could be more easily used with traditional item selection algorithms but has not yet been evaluated with a CAT. In brief, the MP approach consists of replacing the linear predictor of parametric IRT models with an MP (Falk & Cai, 2016a, 2016b; Feuerstahler, 2016; Liang, 2007; Liang & Browne, 2015). Although the approach technically has additional parameters and results in more flexibly estimated IRFs like those from nonparametric approaches, these parameters are not easily interpretable. Thus, the MP approach has been called ‘‘semiparametric’’ or ‘‘quasi-parametric’’ (Liang, 2007; Liang & Browne, 2015). In addition, the MP approach has some distinct practical advantages that we believe permit its further study. In particular, the MP approach can allow calibration of IRFs by maximizing the marginal likelihood or posterior using the expectation-maximization algorithm (EM-MML; Bock & Aitkin, 1981; Mislevy, 1986). This feature is distinct from kernel smoothing (KS; Ramsay, 1991) and smoothed isotonic regression (Lee, 2002, 2007), in which estimation was developed by relying on a proxy of the latent trait that is computed from observed scores. The MP approach is therefore more readily usable in settings where a planned missing data design is used for field testing (i.e., there are missing item responses; Falk, 2019, 2020), in multiple group settings (Falk & Cai, 2016a), and for linking (Feuerstahler, 2019). The MP approach can also be used in conjunction with parametrically modeled items on the same test, which may facilitate a more seamless integration into operational settings. However, simulations studies have also shown that the MP approach is most suitable for large-scale testing as good estimation typically requires larger sample sizes and many items.

Little prior research has studied the performance of flexible or nonparametric IRFs with a CAT. Xu and Douglas (2006) developed two item selection algorithms for the

KS approach based on Kullback–Leibler (K-L; H.-H. Chang & Ying, 1996) information and Shannon entropy (Shannon, 1948). These item selection algorithms avoid the need to compute derivatives of the IRF in order to perform item selection. These authors evaluated the item selection algorithms with a fixed 50-item CAT. Other details of simulations included a 500-item test bank (with true IRFs from a 2PL) calibrated using KS with 1,000 subjects' complete responses. Although the item selection algorithms with KS were found to perform well, the scope of simulations was limited. In particular, KS with both item selection algorithms was compared only with a ''random'' item selection algorithm and comparisons of this approach with traditional IRT models (2PL, 3PL, etc.) was not made. Furthermore, it is unlikely that calibration sample students would be able to complete 500 items each. In later research, Y.-P. Chang et al. (2019) used similar item selection algorithms to evaluate the performance of a nonparametric technique with a cognitive diagnosis model. They found that a nonparametric approach performed better than parametrically estimated IRFs, yet the focus of their study was on smaller sample educational testing contexts.

We therefore have little knowledge of the performance of nonparametric or flexible IRF estimation techniques versus parametric techniques in a CAT under conditions that may be typical of a large-scale test, and no prior research evaluating MP-based models in a CAT. On the one hand, simulations do suggest that flexible IRF estimation techniques such as the MP can improve recovery of the true IRFs, which in turn often leads to better recovery of latent traits (e.g., Falk & Cai, 2016a; Feuerstahler, 2016). However, these previous studies have typically only studied the case where calibration and latent trait estimation is performed on the same sample, and all items on the test are used for all respondents. Use of one calibration sample followed by a CAT using a separate sample is perhaps more similar to cross-validation of the estimated IRFs, and we may not be able to easily extrapolate based on the results of limited previous simulations.

In this article, we present a simulation study that compares the performance of the MP approach, KS, and 2PL in a CAT. We begin by describing each IRF estimation technique, followed by item selection algorithms. Then, we present the method and results of a Monte Carlo simulation study. We finally make concluding remarks.

Method

Studied Item Response Function Estimation Techniques

For the purpose of notation for calibration, consider $i = 1, 2, \ldots, N$ respondents who complete some subset of $j = 1, 2, \ldots, J$ items in the item pool. The 2PL is one of the most commonly used item response models for dichotomous items on both educational and psychological tests. The functional form is essentially that of a logistic regression, in which the item response is regressed on the latent construct, θ. For item j, one way to write the 2PL is as follows:

$$P_j(\theta) = \frac{1}{1 + \exp(-(c_j + a_j\theta))} \tag{1}$$

where c_j is an intercept and a_j is a slope. Conceptually this function traces the probability of a ‘‘correct’’ (or 1; as opposed to an incorrect or 0) response at different levels of the latent trait, θ. The basic idea behind the MP version of this dichotomous item model is to replace the term $a_j\theta$ with a monotonic polynomial, $m_j(\theta)$:

$$P_j(\theta) = \frac{1}{1 + \exp(-(c_j + m_j(\theta)))} \tag{2}$$

Here, the MP is a function of θ:

$$m_j(\theta) = b_{1,j}\theta + b_{2,j}\theta^2 + \cdots + b_{2k_j+1,j}\theta^{2k_j+1} \tag{3}$$

where k_j is a nonnegative integer for item j that controls the order of the polynomial. Due to the form of Equation (2) being a logistic function of $m_j(\theta)$, sometimes this approach is called a ‘‘filtered’’ MP (Feuerstahler, 2016, 2019). Typically coefficients of the polynomial are not directly estimated but are a function of other parameters (e.g., Falk & Cai, 2016a; Feuerstahler, 2016; Liang, 2007). In this way, $P_j(\theta)$ is monotonically increasing in θ but has additional flexibility beyond the 2PL. For both the 2PL and MP approaches, we obtained parameter estimates via EM-MML (e.g., see Falk & Cai, 2016a) with polynomial order selection for the MP done using simulated annealing (Falk, 2019).

KS (Ramsay, 1991) is one of the most popular nonparametric techniques for IRF estimation, possibly due to its availability in programs such as TESTGRAF (Ramsay, 2000) and now in KernSmoothIRT (Mazza et al., 2014). In the context of dichotomously scored items, KS estimated IRFs at any given point along θ resemble a weighted sum of the following form:

$$P_j(\theta) = \sum_{i=1}^{N} w_i(\tilde{\theta}_i)y_{ij} \tag{4}$$

where $w_i(\theta - \tilde{\theta}_i)$ are weights, y_{ij} is examinee i’s response to item j, assuming complete data, and $\tilde{\theta}_i$ is typically a surrogate estimate of the examinee’s score on the latent trait. Values for $\tilde{\theta}_i$ are often chosen based on a (weighted) sum score of all item responses for respondent i. Weights are then often Nadaraya–Watson (Nadaraya, 1964; Watson, 1964) weights as follows,

$$w_i(\theta - \tilde{\theta}_i) = \frac{K\left(\frac{\theta - \tilde{\theta}_i}{h}\right)}{\sum_i K\left(\frac{\theta - \tilde{\theta}_i}{h}\right)} \tag{5}$$

where $K(\cdot)$ is a kernel function and h is a bandwidth. In addition, although Equation (4) appears to be continuous along θ, typically evaluation points along θ are determined, and this may have an impact on scoring and any additional calculations that are performed as part of a CAT. Thus, the grid that is chosen for KS can have implications for how any follow-up scoring or CAT is performed as an item bank would typically need to store such information and would not be able to recompute $P_j(\theta)$ at new θ points on the fly from the original calibration data.

Studied Item Selection Algorithms

In the present article, we consider a fixed-length, item-by-item CAT with $j = 1, 2, \ldots, J$ indexing items in the item pool, $m = 1, 2, \ldots, M$ indexing iterations of the CAT algorithm, and j_m therefore serving as the index of the item administered at iteration m. Then, $\mathbf{y}_{i,j_{m-1}} = [y_{i,j_1} \quad y_{i,j_2} \quad \ldots \quad y_{i,j_{m-1}}]$ are the item responses to administered items prior to iteration m. Let $\tilde{\theta}_{i,m}$ be the score estimate for examinee i at the beginning of iteration m and $\tilde{\theta}_{i,m+1}$ as the updated score estimate after administering item m. At iteration m, item selection algorithms try to choose the next item that will most improve $\tilde{\theta}_{i,m+1}$.

To briefly cover the logic behind MFI (Lord, 1980) and MPWI (van der Linden, 1998), suppose that the true latent trait score for examinee i, θ_i^*, were known and we wished to administer the item(s) that would most reduce the sampling variability of their score estimate, $\hat{\theta}_i$, under maximum likelihood or *expected a posteriori* (EAP; Bock & Mislevy, 1982) scoring, for example. Items that have the highest Fisher information at θ_i^* would be the most optimal to choose. Assuming item parameters implied by a parametric model or the exact shape of the true IRFs were known, Fisher information for dichotomously scored item j is often written as

$$I_j(\theta) = \frac{\left[P_j'(\theta)\right]^2}{P_j(\theta)Q_j(\theta)} \tag{6}$$

where $P_j'(\theta) = \frac{\partial P_j(\theta)}{\partial \theta}$ and $Q_j(\theta) = 1 - P_j(\theta)$. Note that this expression is generally given when defining expected Fisher information, yet for exponential family item response models such as those in Equations (1) and (2) is equivalent to that of observed information. Note also that this requires partial derivatives of $P_j(\theta)$ with respect to θ, which are easily obtained for the 2PL and MP approaches but not in general with KS. As θ_i^* is unknown, MFI uses the current interim estimate $\hat{\theta}_{i,m}$ as a substitute for θ in Equation (6) and then the top item is chosen.

Due in part to instability in $\hat{\theta}_{i,m}$, especially in the early stages of a CAT, the item chosen under MFI may not be optimal for the examinee. MPWI quantifies uncertainty in $\hat{\theta}_{i,m}$ by computing information for each item integrating across the posterior distribution for $\hat{\theta}_{i,m}$. Let $\pi(\theta)$ be the prior density and

$$L(\theta|\mathbf{y}_{i,j_{m-1}}) = \prod_{k=1}^{m-1} P_{j_k}(\theta)^{y_{i,j_k}} Q_{j_k}(\theta)^{1-y_{i,j_k}} \tag{7}$$

be the likelihood for examinee i at the start of iteration m. Then, the posterior weighted information is defined as follows (see also Magis & Raiche, 2012):

$$PWI_j(\theta) = \int I_j(\theta) L(\theta | \mathbf{y}_{i,j_{m-1}}) \pi(\theta) d\theta \tag{8}$$

where the integral may be computed using numerical methods (e.g., rectangular quadrature). MPWI then proceeds at any given iteration by computing PWI_j for all items under the current posterior, and choosing the item with the largest value.

Finally, K-L information (H.-H. Chang & Ying, 1996) resembles the likelihood ratio between the unknown but true value θ_i^*, and some other value, θ:

$$K_j(\theta_i^* || \theta) = E_{\theta_i^*} \log \left[\frac{L_j(\theta_i^*; Y_{ij})}{L_j(\theta; Y_{ij})} \right] \tag{9}$$

with the expectation over possible responses for the item (values for the random variable Y_{ij}), which in this case is $L_j(\theta; Y_{ij}) = P_j(\theta)^{Y_{ij}} Q_j(\theta)^{1-Y_{ij}}$. Conceptually, K-L determines which newly administered item would allow us to tell the difference between θ_i^* and some other value on the latent scale using a likelihood ratio.

While MFI and MPWI computations for each item rely on some estimate or posterior for θ_i^*, computation of K-L in Equation (9) relies on a stand-in for both θ_i^* and some other possible value(s) for θ. To achieve this, K-L information is computed using the current interim estimate $\hat{\theta}_{i,m}$ for θ_i^*, and a range of values for θ in the vicinity of $\hat{\theta}_{i,m}$ are considered via an integral:

$$K_j(\hat{\theta}_{i,m}) = \int_{\hat{\theta}_{i,m} - \delta_m}^{\hat{\theta}_{i,m} + \delta_m} K_j(\hat{\theta}_{i,m} || \theta) d\theta \tag{10}$$

with a typical choice for $\delta_m = c/\sqrt{m}$. In simulations, $c = 3$ is often chosen (e.g., H.- H. Chang & Ying, 1996) though could be chosen to determine the range of integration corresponding to a desired coverage probability for θ_i^* based on an interval around $\hat{\theta}_m$. The term δ_m thus controls the range of integration and is wide in the early stages of the CAT, to effectively obtain a ‘‘global’’ index of information. At later stages of the CAT (i.e., a larger value for m), the range of integration decreases and Equation (10) would become more local, properly reflecting better information regarding the latent trait estimate.

Note that computation of Equation (10) does not require any derivative computations, only the ability to evaluate $P_j(\theta)$ at particular values of θ. Thus, it is well-suited for use also with KS (in addition to 2PL and MP) approaches to IRF estimation. However, evaluation of the integral may require numerical integration, which still requires choosing quadrature nodes or grid points along θ. These nodes will typically not match the exact evaluation points under KS obtained during calibration. Thus, to obtain the values of the IRF under KS at each quadrature node to compute Equation (10), we used linear interpolation based on the already estimated IRF. We suppose this is the most likely situation if KS were to be used to in an item bank.

Study Design and Simulation Details

For clarity, we separate the study design details into two sections. First, we describe how calibration data were simulated and how models were estimated. On the basis of IRF estimation from calibration data, we describe how CAT simulations were then conducted.

Calibration and Data Generating Models. To attempt to construct realistic field testing conditions, we generated calibration sample data under two broad conditions: Complete data and missing data. Within each of these two conditions, we also varied sample size and the percentage of items that had an IRF that departed drastically from the traditional 2PL model (nonstandard items). This resulted in a total of eight different data generating conditions for the calibration samples.

For the complete data conditions, all respondents completed all items. The number of items was fixed at $J = 100$, as we expect this to be approximately the largest amount of items that respondents may reasonably complete, especially if such items represent educational test items. The number of respondents ($N = 1,000$ or $N = 3,000$) was fully crossed with the percentage of nonstandard items (30% or 70%).

For the missing data conditions, each respondent completed only a random subset of 40 items, which is similar to some recent large-scale educational tests (e.g., Smarter Balanced Assessment Consortium, 2017). Since the number of items under this condition was fixed at $J = 200$, this represents 80% missing data. To compensate for missing data, we would expect research teams under similar testing conditions to utilize more respondents. Thus, the sample size conditions were larger. Sample size ($N = 5,000$ or $N = 10,000$) was again crossed with the percentage of nonstandard items (30% or 70%).

All items were dichotomous, and standard items were generated using a normal cumulative distribution function (CDF) as the IRF, $P_j(\theta) = \Phi(\theta|\mu_{1j}, \sigma^2_{1j})$, with $\mu_1 \sim \text{unif}(-2.75, 2.75)$ and $\sigma_1 \sim \mathcal{N}(2, .4^2)$ drawn randomly across items. Although this does not strictly follow the 2PL, it may be reconceived as following a normal ogive model for which the 2PL provides a very close approximation. Nonstandard items were generated with the following mixture of normal CDFs: $P_j(\theta) = \pi_1\Phi(\theta|\mu_{1j}, \sigma^2_{1j}) + \pi_2\Phi(\theta|\mu_{2j}, \sigma^2_{2j}) + \pi_3\Phi(\theta|\mu_{3j}, \sigma^2_{3j})$. Proportions were generated randomly across items, $\pi_1 \sim \text{unif}(.1, .6), \pi_2 \sim \text{unif}(.1, .3), \pi_3 = 1 - \pi_1 - \pi_2$, as were standard deviations, $\sigma \sim \mathcal{N}(.7, .2^2)$ for $\sigma_1, \sigma_2, \sigma_3$ with any values less than .2 winsorized at .2. To provide variation in overall difficulty, means of the CDFs were pieced together in the following way: $\mu_1 = \mu + \delta_1, \mu_2 = \mu + \delta_2, \mu_3 = \mu + \delta_3$. Thus, $\mu \sim \text{unif}(-2.5, 2.5)$ provided some overall control of difficulty, and the remaining parameters controlled the center of the CDFs, $\delta_1 \sim \mathcal{N}(-2.2, .2^2)$, $\delta_2 \sim \mathcal{N}(2.2, .2^2)$, $\delta_3 \sim \mathcal{N}(0, .2^2)$.

For each data generating condition, a single calibration sample was generated, with a standard normal θ assumed. Although it may be preferable to have several calibration samples per data generation condition, and then repeatedly conduct CAT simulations with each calibration sample, such an approach is still very computationally intensive if conducting thousands of CAT simulations for each calibration sample. We therefore decided against this approach and comment further on this issue in the Discussion section.

To each calibration sample, several models were fit to obtain IRF estimates for later use in CAT simulations. The exact models depended on the data generating conditions. The KS approach using KernSmoothIRT (Mazza et al., 2014) with default settings was utilized for all complete data generating conditions. As noted, this method is possible but not ideal when there are missing item responses and so was used with only complete data conditions. To obtain θ_i for smoothing, the default behavior is to compute sum scores (equal weight for each item), obtain percentile ranks (with ties broken in order of appearance), and then use a normalizing transformation using quantiles of a standard normal distribution. Although multiple choices for a kernel function are possible, a Gaussian kernel is default and a reasonable choice. The bandwidth defaults to the so-called Silverman rule of thumb (Silverman, 1986), which was .266 when $N = 1,000$ and .214 when $N = 3,000$. And finally, the evaluation grid defaults to 51 equally spaced points between ± 3.09 when $N = 1,000$ and ± 3.40 when $N = 3,000$.

In addition, a 2PL model was fit to all data sets. Finally, an MP-based model was fit to each data set. For the MP models, the order of the polynomial was determined through use of simulated annealing as described by Falk (2019) with up to $k = 3$ considered, a starting temperature of 5, a logarithmic temperature schedule as described by Stander and Silverman (1994), and using Akaike information criterion as the optimization criterion. The number of iterations was set to 800 for complete data and 1,600 for missing data since the latter had a larger item bank and may require more iterations to find a good solution. Simulated annealing was used as not all items follow a nonstandard IRF and a relatively large number of items is not easily amenable to a step-wise approach for selecting polynomial order for each item. *OpenMx* and *rpf* packages were used for fitting the 2PL and MP approaches with custom R code for simulated annealing (Neale et al., 2016; Pritikin, 2016; R Core Team, 2017). For both 2PL and MP models, EM-MML was used with integrals evaluated by rectangular quadrature with 101 equally spaced nodes on ± 5, and M-step and E-step tolerance for convergence was 1.0×10^{-9} and 1.0×10^{-7}, respectively.[1]

Computer Adaptive Test Simulations. We first describe simulation conditions that were fixed across all CAT simulations, followed by manipulated factors. First, θ for simulees was generated in one of two ways: from a standard normal distribution and at discrete points along θ (-2 to 2 in 0.5 increments). Under each of these conditions and each manipulated condition described below, 1,000, simulees were generated and the true IRFs were used to generate their hypothetical item responses. Such a data generation technique allowed us to tell whether overall some IRF estimation techniques resulted in better recovery of θ as well as whether there were certain locations along θ where recovery was better/worse. For all simulees, a fixed CAT length of 25 items was utilized, with interim and final θ estimates using the EAP scoring method. Depending on the item selection algorithm, a starting $\hat{\theta}$ of 0 (MFI and K-L), or a starting prior of a standard normal distribution (MPWI), was used. To make the simulations as ‘‘fair’’ as possible for KS, the grid points used for both interim EAP scoring (with MFI and K-L) and for representing the posterior distribution under MPWI were chosen as the same 51 grid points used for KS estimation as defined in the Calibration and Data Generating Models section. Programming for CAT simulations was based on custom R code with analytical derivatives for the MP and 2PL provided by *rpf.*

Manipulated factors for CAT simulations involved (1) the source of IRF estimation, and (2) the item selection algorithm. The IRFs used in the CAT involved up to four conditions: those from the calibration samples (2PL, KS for complete data only, and the MP approach), as well as use of the true IRF. Use of the 2PL allowed us to gauge whether KS or MP have much of an advantage over a parametric model and use of the true IRF allows a benchmark for the best possible IRF that could be used. Three possible item selection algorithms were crossed with the available IRF estimation techniques (where possible): MFI and MPWI (for the 2PL and MP only) and K-L information.[2]

Results

In what follows, we first briefly present results pertaining to the IRF recovery for the eight calibration samples. These results are presented to frame understanding of how different IRF estimation techniques may recover true IRFs, which may indirectly affect CAT performance. Following such results, we will turn to the primary results of interest regarding performance of each type of IRF and item selection method in CAT simulations. As the amount of data collected for the study is vast, this represents our best attempt at understanding the pattern of results.

Calibration Results

Recovery of IRFs was assessed using root integrated mean square error (RIMSE), which is computed as a squared discrepancy between the true, $P_j(\theta)$, and estimated IRF, $\hat{P}_j(\theta)$, integrated across the latent distribution with the square root taken of the final quantity. Here the integral was approximated using rectangular quadrature with $Q = 101$ nodes between -5 and 5:

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

where $\phi(\cdot)$ is the standard normal density function. For KS items, the sum was taken over the evaluation points defined by the KS model. RIMSE was computed for each individual item using a standard normal latent trait distribution. Averaging over items within each cell of the design, MP tended to outperform the 2PL under most

Table 1. Mean RIMSE for Item Banks Under Complete Data Collection Design.

		Model		
Sample size (*N*)	Proportion nonstandard (%)	2PL	MP	KS
1,000				
	30	.07	.07	.10
	70	.11	.10	.11
3,000				
	30	.04	.03	.04
	70	.09	.04	.05

Note. RIMSE = root integrated mean square error; 2PL = two-parameter logistic; MP = monotonic polynomial; KS = kernel smoothing.

Table 2. Mean RIMSE for Item Banks Under Missing Data Collection Design.

		Model	
Sample size (*N*)	Proportion nonstandard (%)	2PL	MP
5,000			
	30	.07	.09
	70	.12	.09
10,000			
	30	.05	.05
	70	.09	.07

Note. RIMSE = root integrated mean square error; PL = parameter logistic; 2PL = two-parameter logistic; MP = monotonic polynomial.

conditions (Tables 1 and 2). This was especially true when there was a larger percentage of nonstandard items (70%) or a larger sample size (*N* = 3,000 with complete data or *N* = 10,000 under missing data). Otherwise, the MP and 2PL performed similarly, and the 2PL performed better under one condition (*N* = 5000, missing data, 30% nonstandard items). Recall KS was used in complete data conditions, yet performed worse than the MP under all these conditions and equal to or worse than the 2PL in all but the *N* = 3,000, 70% nonstandard item condition.

Examining the distribution for RIMSE for standard and nonstandard items separately, it was apparent that gains with the MP were due mainly to better estimation for nonstandard items (Figure 1). The MP did not perform better than the 2PL for standard items, and sometimes performed slightly worse. The KS approach performed on par with MP for nonstandard items but clearly worse than the 2PL and MP approaches for standard items.

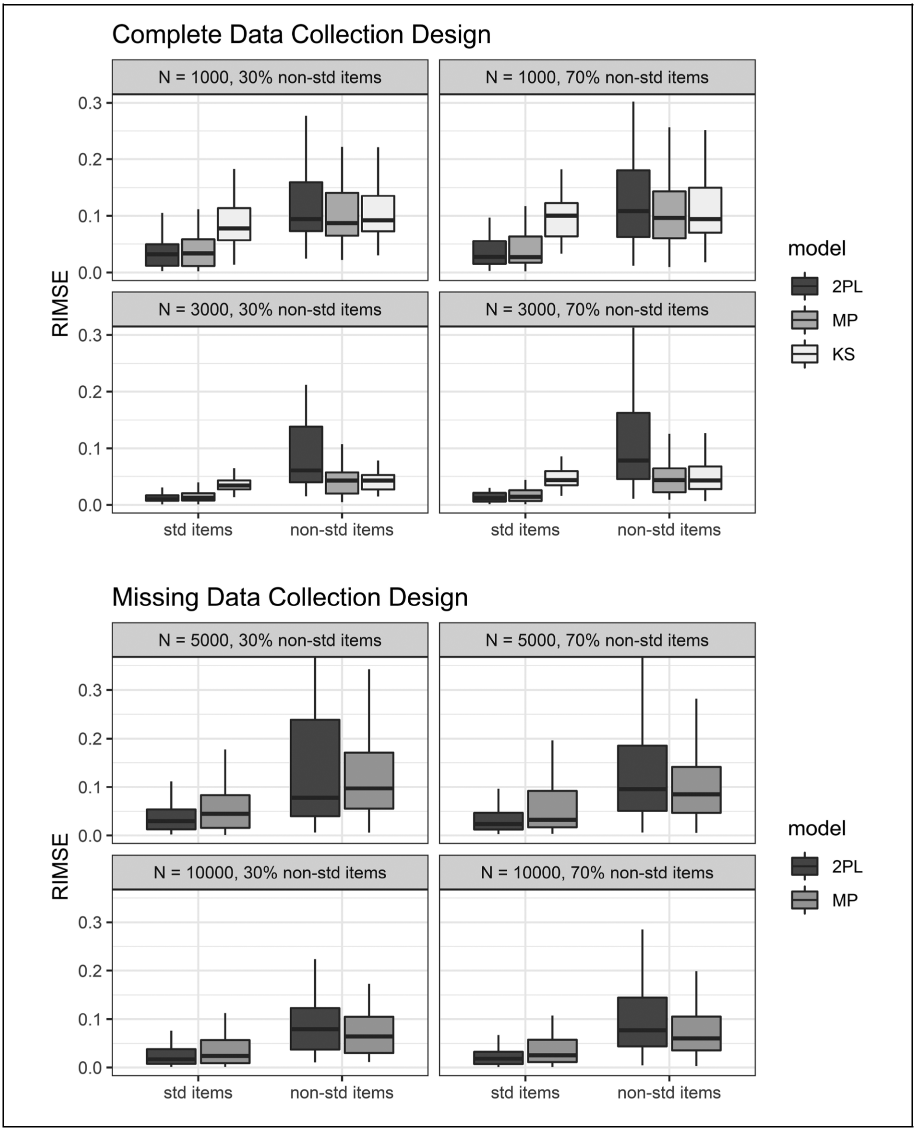

Figure 1. Recovery of response functions for standard (std) and nonstandard (nonstd) items for each calibration.

We also compared the recovery of item-level information among the estimated 2PL and MP items. As shown in the Supplemental Material, available online, MP and 2PL items provided equally accurate information, although nonstandard items tended to have worse information accuracy than standard items when fit to both models.

Table 3. Average RMSE of Latent Trait Scores for CAT Simulations Under Complete Data Calibration, Standard Normal Latent Traits.

			Model			
Item selection	Sample size (*N*)	Proportion nonstandard (%)	True	2PL	MP	KS
KL	1,000					
		30	.363	.381	**.379**	.386
		70	.341	.385	.384	**.378**
	3,000					
		30	.375	.396	.396	**.395**
		70	.360	.412	**.378**	.380
MFI	1,000					
		30	.359	.383	**.377**	
		70	.343	.384	**.380**	
	3,000					
		30	.373	.396	**.395**	
		70	.361	.412	**.361**	
MPWI	1,000					
		30	.361	.383	**.375**	
		70	.349	**.384**	.390	
	3,000					
		30	.375	.398	**.395**	
		70	.364	.407	**.375**	

Note. Excluding the true model condition, the best performing method in each row appears in bold. Sample size refers to that used in calibration. RMSE = root mean square error; KL = Kullback–Leibler information; MFI = maximum Fisher information; MPWI = maximum posterior weighted information; True = true model; 2PL = two-parameter logistic; MP = monotonic polynomial; KS = kernel smoothing.

Computer Adaptive Test Simulations

The primary outcome of interest from CAT simulations was recovery of true latent trait scores. For this purpose, we examined root mean square error: $RMSE = \left(N^{-1}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2\right)^{1/2} \times 100$, where $\hat{\theta}_i$ is the estimated latent trait score and θ_i is the true latent trait for simulee i. When latent traits were drawn from a standard normal distribution and when complete data was available for calibration, both MP and KS tended to outperform the 2PL approach, regardless of the item selection algorithm (Table 3). Under complete data calibration, for example, the 2PL only led to lower average RMSEs than the MP in a single cell: MPWI item selection with $N = 1,000$ and 30% nonstandard items.

When the data calibration had missing data, the MP and 2PL often led to similar average RMSEs as each other (Table 4). MP always led to equal or lower average RMSEs with MFI item selection, but these patterns were more mixed with KL and MPWI item selection. As an example with KL and MPWI, with 70% nonstandard items and $N = 5,000$, MP had average RMSE that was $\geq .025$ better than the 2PL.

Table 4. Average RMSE of Latent Trait Scores for CAT Simulations Under Missing Data Calibration andStandard Normal Latent Traits.

			Model		
Item selection	Sample size (*N*)	Proportion nonstandard (%)	True	2PL	MP
KL	5,000				
		30	.331	**.370**	.374
		70	.286	.361	**.335**
	10,000				
		30	.327	.352	**.350**
		70	.311	**.347**	.350
MFI	5,000				
		30	.318	.373	**.372**
		70	.284	.364	**.339**
	10,000				
		30	.327	.351	**.348**
		70	.310	**.347**	**.347**
MPWI	5,000				
		30	.331	.369	**.367**
		70	.288	.360	**.329**
	10,000				
		30	.327	.350	**.343**
		70	.312	**.346**	.350

Note. Excluding the true model condition, the best performing method in each row appears in bold. Sample size refers to that used in calibration. RMSE = root mean square error; KL = Kullback–Leibler information; MFI = maximum Fisher information; MPWI = maximum posterior weighted information; True = true model; 2PL = two-parameter logistic; MP = monotonic polynomial; KS = kernel smoothing.

With 70% nonstandard items and $N = 10,000$, the 2PL was better but by a much smaller amount ($\leq .004$). In summary, based on examination average performance with normally distributed latent traits, use of MP calibrated items performed as well as or better than the 2PL and similarly to KS.

Turning to RMSE at discrete points along θ, we concentrate primarily on results using KL item selection as this allows comparison with KS estimated IRFs (Figures 2 and 3).[3] Based on such results, it is clear that any performance advantage of one method of IRF estimation versus another is not necessarily consistent across the entire range of the latent trait. For example, the MP had an advantage over the 2PL for much of the latent trait for conditions with a larger proportion of nonstandard items (70%), yet in some regions of the latent trait (especially near $\theta = .5$ or 1), these differences were small or the 2PL outperformed the MP. Differences among approaches with 30% nonstandard items were more difficult to visualize, thus indicating similar performance, except perhaps for the KS performing slightly worse than other methods in the middle of the distribution when $N = 1,000$. It is also possible that these pattern of results may vary across calibrated item banks.

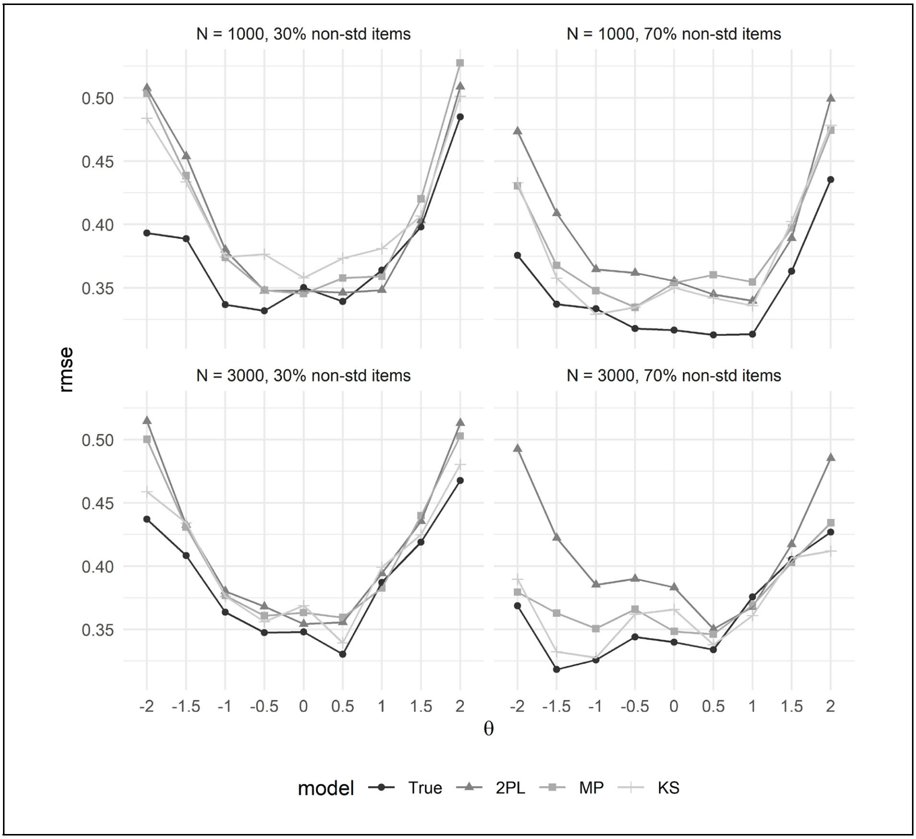

Figure 2. RMSE of latent trait scores at discrete points along θ with KL item selection, complete data calibration.

Note. KL = Kullback–Leibler information; RMSE = root mean square error; non-std = nonstandard.

Discussion

The presented simulation study examined the performance of MP and KS IRF estimation techniques for use with a CAT, and compared them with a standard 2PL approach using item selection techniques based on KL information, MFI, and MPWI. Our results demonstrate that MP and KS approaches lead to comparable or better latent trait recovery than the 2PL.

Despite the promise of the MP and KS approaches, it is difficult to pinpoint exact conditions under which such an approach is universally preferable to standard approaches such as the 2PL. In retrospect, different IRFs can still result in very similar latent trait estimates (e.g., Yen, 1981). More substantial departures from the 2PL in the form of more extreme IRFs (including nonmonotonic) may need to be present in the item bank for the various methods to perform much more differently than one

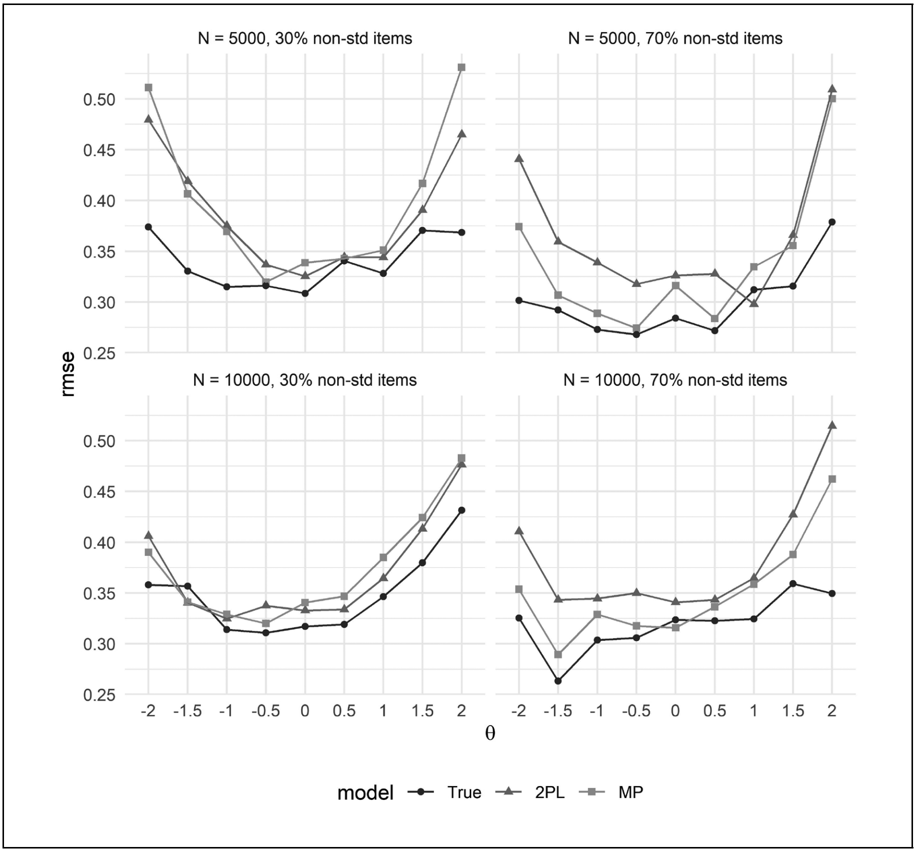

Figure 3. RMSE of latent trait scores at discrete points along θ with KL item selection, missing data calibration.

Note. KL = Kullback–Leibler information; RMSE = root mean square error; non-std = nonstandard.

another. Our manipulated conditions mainly varied features of the calibration sample to mimic conditions that might be used for a large test (item banks of 100 or 200, and complete or planned missing data collection). While prior simulation studies have found that MP and KS can on average recover IRFs better than the 2PL when nonstandard items are in an item bank (e.g., Falk & Cai, 2016a; Feuerstahler, 2016), in any given calibration sample it may be that such gains do not always clearly materialize or do not then lead to subsequent gains in scoring or CAT performance. With some exceptions, the MP tended to perform better with larger calibration samples and with a larger proportion of nonstandard items. It is suggested that future research may focus on features of the CAT itself (test length, stopping criteria) that may affect performance, as well as on identifying the conditions for which flexible IRF estimation provides a clear advantage over the standard approaches. Our study utilized a realistic calibration phase prior to conducting CAT simulations. We thought this necessary for studying the relative performance of IRF estimation techniques in a CAT. However, this approach also makes it slightly difficult to know whether the relative performance observed in any given cell of the CAT simulation design was due in part to random sampling fluctuation from doing only a single calibration per cell. Although this issue could be addressed by doing multiple calibrations per cell and then multiple CAT simulations, such analyses demand a large amount of computational time and space. In addition, given the size of the item banks (100 and 200 items, depending on the condition) relative to the length of the CAT (25 items), we expected that doing only a single calibration would still be informative.

This study was also apparently the first to utilize a flexible IRF estimation technique (the MP) in conjunction with item selection algorithms that require analytical derivatives and are often used in operational settings (MFI and MPWI). We found little difference among item selection algorithms,[4] though we did not present any prior theory to favor any particular technique. This result holds promise for operational programs that may consider a nonparametric or semiparametric approach to IRF estimation but may prefer to use familiar, derivative-based item selection algorithms or would prefer to implement changes in stages to better ensure quality control.

In closing, we believe that the MP approach may be particularly well-suited to applications in CAT because it allows for both flexibly estimated IRFs and analytic derivatives. The initial results presented in this article indicate that the MP approach estimates latent traits as well or better than a standard approach when a significant proportion of nonstandard items exist and in the context of a planned missing data field test design.

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 research, authorship, and/or publication of this article: We acknowledge the support of the Natural Science and Engineering Research Council of Canada (Funding Reference Number RGPIN-2018-05357 and DGECR-2018-00083).

ORCID iDs

Carl F. Falk https://orcid.org/0000-0002-4788-7206
Leah M. Feuerstahler https://orcid.org/0000-0002-7001-8519

Supplemental Material

The online supplemental materials are available at http://journals.sagepub.com/doi/suppl/10.1177/00131644211014261.

Notes

1. The code base that enables this estimation approach is now available as an R package on GitHub: https://github.com/falkcarl/mpirt.git
2. K-L information that additionally weighted by the likelihood or by the posterior (similar to that implemented in the *catR* package) was also attempted but did not appear to result in any better performance than the K-L results we report here.
3. Results for other item selection methods are available in Supplemental Material, available online.
4. Though we mention in Note 2 that KL item selection weighting across the likelihood and posterior, as implemented in *catR* (Magis & Raiche, 2012), does not perform very well.

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. https://doi.org/10.1007/BF02293801

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444. https://doi.org/ 10.1177/014662168200600405

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*(3), 213-229. https://doi.org/10.1177/014662169602000303

Chang, Y.-P., Chiu, C.-Y., & Tsai, R.-C. (2019). Nonparametric CAT for CD in educational settings with small samples. *Applied Psychological Measurement*, *43*(7), 543-561. https://doi.org/10.1177/0146621618813113

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* (pp. 75-85). Springer Nature. https://doi.org/10.1007/978-3-030-01310-3_7

Falk, C. F. (2020). The monotonic polynomial graded response model: Implementation and a comparative study. *Applied Psychological Measurement*, *44*(6), 465-481. https://doi.org/10.1177/0146621620909897

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*(2), 434-460. https://doi.org/10.1007/s11336-014-9428-7

Falk, C. F., & Cai, L. (2016b). Semi-parametric item response functions in the context of guessing. *Journal of Educational Measurement*, *53*(2), 229-247. https://doi.org/10.1111/jedm.12111

Feuerstahler, L. M. (2016). *Exploring alternate latent trait metrics with the filtered monotonic polynomial IRT model* [Unpublished doctoral dissertation]. University of Minnesota.

Feuerstahler, L. M. (2019). Metric transformations and the filtered monotonic polynomial item response model. *Psychometrika*, *84*(1), 105-123. https://doi.org/10.1007/s11336-018-9642-9

Lee, Y.-S. (2002). *Applications of isotonic regression in item response theory* [Unpublished doctoral dissertation] University of Wisconsin–Madison.

Lee, Y.-S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement*, *31*(2), 121-134. https://doi.org/10.1177/0146621606290248

Liang, L. (2007). *A semi-parametric approach to estimating item response functions* [Unpublished doctoral dissertation]. 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*(1), 5-34. https://doi.org/10.3102/1076998614556816

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.

Magis, D., & Raiche, G. (2012). lavaan: Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48*(8), 1-31. https://doi.org/10.18637/jss.v048.i08

Mazza, A., Punzo, A., & McGuire, B. (2014). KernSmoothIRT: An R package for kernel smoothing in item response theory. *Journal of Statistical Software*, *58*(6), 1-34. https://doi.org/10.18637/jss.v058.i06

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*(2), 177-195. https://doi.org/10.1007/BF02293979

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Nadaraya, E. A. (1964). On estimating regression. *Probability Theory and Its Applications*, *9*(1), 141-142. https://doi.org/10.1137/1109020

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kickpatrick, R. M., Estabrook, R., Bates, T. C., Maes, H. H., & Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*(2), 535-549. https://doi.org/10.1007/s11336-014-9435-8

Pritikin, J. N. (2016). *rpf: Response probability functions* R package version 0.53) [Computer software]. R-project.org. https://CRAN.R-project.org/package=rpf

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*(4), 611-630. https://doi.org/10.1007/BF02294494

Ramsay, J. O. (2000). *TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data* [Computer software]. TestGraf. https://www.psych.mcgill.ca/misc/fda/downloads/testgraf/TestGraf98.pdf

R Core Team. (2017). *R: A language and environment for statistical computing*. R-project.org. http://www.R-project.org

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall.

Smarter Balanced Assessment Consortium. (2017). *Smarter Balanced Assessment Consortium: 2016-17 technical report*. Smarter Balanced. https://portal.smarterbalanced.org/library/en/2016-17-summative-assessment-technical-report.pdf

Stander, J., & Silverman, B. W. (1994). Temperature schedules for simulated annealing. *Statistics and Computing*, *4*(1), 21-32. https://doi.org/10.1007/BF00143921

Stout, W. (2001). Nonparametric item response theory: A maturing and applicable measurement modeling approach. *Applied Psychological Measurement*, *25*(3), 300-306. https://doi.org/10.1177/01466210122032109

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*(2), 201-216. https://doi.org/10.1007/BF02294775

Watson, G. S. (1964). Smooth regression analysis. *Sankhya:The Indian Journal of Statistics, Series A*, *26*(4), 359-372.

Xu, X., & Douglas, J. A. (2006). Computerized adaptive testing under nonparametric IRT models. *Psychometrika*, *71*(1), 121-137. https://doi.org/10.1007/s11336-003-1154-5

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 245-262. https://doi.org/10.1177/014662168100500212