Model Selection for Monotonic Polynomial Item Response Models

Carl F. Falk

Abstract One flexible approach for item response modeling involves use of a monotonic polynomial in place of the linear predictor for commonly used parametric item response models. Since polynomial order may vary across items, model selection can be difficult. For polynomial orders greater than one, the number of possible order combinations increases exponentially with test length. I reframe this issue as a combinatorial optimization problem and apply an algorithm known as simulated annealing to aid in finding a suitable model. Simulated annealing resembles Metropolis-Hastings: A random perturbation of polynomial order for some item is generated and acceptance depends on the change in model fit and the current algorithm state. Simulations suggest that this approach is often a feasible way to select a better fitting model.

Keywords Combinatorial optimization · Nonparametric item response theory · Monotonic polynomial · Balanced incomplete block design

Many standard unidimensional item response models assume a normally distributed latent trait and a simplistic relationship between the latent trait and the item responses. For example, the two-parameter logistic model (2PL) represents a multivariate extension of logistic regression, where the log-odds of obtaining a correct response to the items is a linear function of the latent trait (Birnbaum, 1968). This relationship may not be expected to hold for all educational and psychological constructs (Meijer & Baneke, 2004), and violations may arise from population heterogeneity in exposure to unique item content (Falk & Cai, 2016b) or items that require multiple steps in order to complete (Lee & Bolt, 2018). Additional flexibility in the trait-response relationship is possible, including but not limited to nonparametric Kernel smoothing (Ramsay, 1991), smoothed isotonic regression (Lee, 2007), Bayesian nonparametric techniques (Duncan & MacEachern, 2013), normal ogive models that assume

C. F. Falk (✉)
Department of Psychology, McGill College, McGill University, 7th Floor, Montreal, QC 2001, H3A 1G1, Canada
e-mail: carl.falk@mcgill.ca
URL: https://www.mcgill.ca/psychology/carl-f-falk

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_7

heteroscedastic errors (Molenaar, 2015), and splines (Ramsay & Wiberg, 2017). Alternatively, if the source of this assumption violation stems in part from a non-normal trait distribution, one could directly model such non-normality (Woods, 2007).

The focus of this paper is on a monotonic polynomial (MP) approach to flexible item response function (IRF) estimation (Falk & Cai, 2016a, 2016b; Liang & Browne, 2015). The basic idea behind MP item response models is to replace the linear predictor of a standard item response model with an MP. Like nonparametric techniques, MP models make few assumptions about the underlying process that produces nonstandard response functions. Rather, increasing polynomial order allows MP models to approximate many different functional forms, regardless of whether the MP is the true model (Feuerstahler, 2016). In contrast to the 2PL, a logistic function of a monotonic polynomial models the log-odds of a correct response as a polynomial function of the latent trait with constraints imposed such that this relationship is monotonic.

We believe the MP approach warrants further study for its potential to fulfill several needs of large scale or operational testing. For example, a psychometrician may use an MP-based model to improve item fit for a few items on a long test, allowing retention of expensive-to-develop items, but still use a traditional item model such as the 2PL or three-parameter logistic (3PL) for the remaining test items. Since MP-based models can also be explained using an analogy with polynomial regression, MP-based approaches may be more substantively interpretable to some stakeholders. We also conjecture that the derivatives necessary for MP-based item models to be used in a computer adaptive test with traditional item selection strategies are readily available in closed form, in contrast to some other approaches (Xu & Douglas, 2006). Finally, a testing program that has hundreds of items is likely to employ a planned missing data design. It would otherwise be burdensome to expect respondents to complete all such test items in a diligent manner. MP-based item models can be used in conjunction with maximum marginal likelihood (MML) estimation (Bock & Aitkin, 1981), which can be used with planned missing data designs and investigations of differential item functioning (Falk & Cai, 2016a).

1 The Computational Problem

One potential barrier for MP-based models involves a computational problem in selecting polynomial order. To further understand, consider the IRF for a logistic function of a monotonic polynomial (Falk & Cai, 2016a; Liang & Browne, 2015):

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

where $j = 1, \ldots, n$ indexes n test items, θ corresponds to the latent trait, and $m_j(\theta)$ is a polynomial term:

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

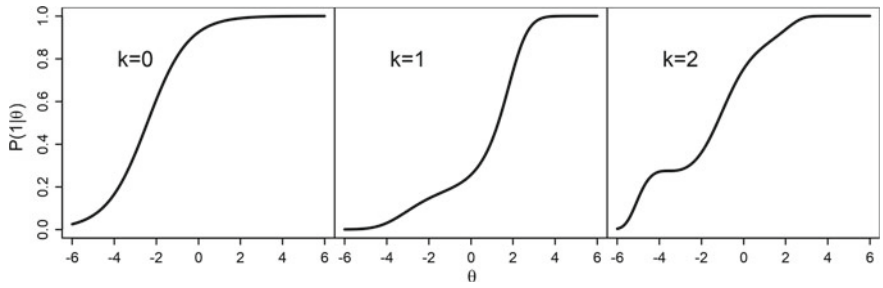

Fig. 1 Example response functions for three different order polynomials

$m_j(\theta)$ is parameterized to be monotonic increasing and has a non-negative derivative with respect to θ. This is accomplished in part by a polynomial with an odd number of terms: $2k_j + 1$, where k_j is a non-negative integer that controls polynomial order for item j (see Fig. 1). In addition, the coefficients, $b_{1,j}, b_{2,j}, \ldots, b_{2k_j+1,j}$, are not directly estimated, but are a function of $2k_j + 1$ other parameters with constraints that maintain monotonicity. Other MP models have been developed based on the 3PL, generalized partial credit, and graded response models (Falk, 2018; Falk & Cai, 2016a, 2016b). When $k_j = 0$, these models reduce to their standard counterparts (e.g., Eq. 1 reduces to the 2PL).

The key to the computational problem concerns the selection of k_j, which may be different for each item. This problem is a byproduct of using MML for estimation: Selection of k_j for one item may affect item fit for other items and overall model fit. In one investigation, Falk and Cai (2016a) employed a step-wise approach whereby AIC was used to select a single increase in polynomial order for one item at a time. This approach is difficult to use with a long test as each step would require fitting n models. For example, if $n = 100$, then 100 models must be fit before increasing polynomial order for a single item. In a different paper, Falk and Cai (2016b) experimented with use of summed score item fit statistics, $S - X^2$ (Orlando & Thissen, 2000), to screen for items that may be good candidates for use of an MP. Although this approach arguably improved fit, $S - X^2$ had power that was less than expected to detect non-standard items, and using summed score based item fit statistics may not always be desirable with missing data. If an observed score substitute for θ is used in estimation instead, then the modeler may proceed item by item in selection of k_j. However, this approach may not readily handle multiple group models or models with missing data.

1.1 A Possible Solution

We reframe the selection of k_j for each item as a combinatorial optimization problem. If we consider k_j for each item from 0 to 2, then there are 3^n possible combinations of

polynomial order. Clearly for large n, there may be many combinations and therefore too many possible models to actually try out even with a modern computer. Further suppose that there is some combination of polynomial order that may be optimal (e.g., according to information criterion such as AIC or BIC). In addition to a stepwise approach being computationally slow, it may also be prone to getting stuck at a local optimum.

Although there are a number of combinatorial optimization algorithms suitable for finding an approximate global optimum, we chose to experiment with simulated annealing (Černý, 1985; Kirkpatrick, Gelatt, & Vecchi, 1983), which has seen some use by psychometricians (Drezner & Marcoulides, 1999; Edwards, Flora, & Thissen, 2012). SA gets its name in part from an analogy to metallurgy, yet we find it more intuitive to explain its workings by analogy to Metropolis-Hastings (MH). Given some model, M_s, at iteration s, SA has the following steps:

1. Generate some candidate model, M_s^*, from a list of possible neighboring models in a well-defined search space.
2. Compute *energy* for the candidate, $e(M_s^*)$, and current model, $e(M_s)$.
3. Determine acceptance/rejection of the candidate, M_s^*, based on the difference in energy, $e(M_s^*) - e(M_s)$, and the current *temperature*, t_s, which represents the current algorithm state.
4. Repeat 1–3 and stop based on an iteration maximum, S, or termination rule.

In the same way that MH will tend to move towards and sample from high-density regions of a probability distribution, SA will tend to move towards and select models in regions of a search space that have better fit. In our application, we allowed values for k_j between 0 and 2, which defines the search space as the 3^n possible polynomial order combinations. We considered a neighboring model to be a random increment or decrement of 1 to k_j for one or two items that were randomly sampled with uniform probability. For example, if item 5 were to be randomly selected and the current $k_5 = 1$, then the candidate could only change to $k_5 = 0$ or $k_5 = 2$ (selected with equal probability). If $k_5 = 0$, then the candidate had $k_5 = 1$. k_j for all other items would remain as-is. Changing k_j by only one at a time for each item and only for a couple of items may allow a reduction in estimation difficulty. For example, use of parameter estimates from a lower-order polynomial may be used as starting values for some parameters when estimating models with higher-order polynomials. However, defining neighbors and the search space in this way, it is possible to move from one state of the search space (e.g., all $k_j = 0$) to the furthest state (e.g., all $k_j = 2$) within only 300 or 150 steps or less if $n = 100$ and either one or two items' polynomials are perturbed at each step.

Energy is a function of the fitted model and defines its optimality. For this purpose, we used $e(\cdot)$ to calculate either AIC or BIC. Thus, lower energy indicates better fit. The acceptance probability of M_s^* was based on the following,

$$\min\left\{1, \exp(-(e^* - e)/t_s)\right\} \tag{3}$$

where we use e^* and e as shorthand for $e(M_s^*)$ and $e(M_s)$, respectively. In other words, if M_s^* has lower energy (or improves model fit), it is accepted with certainty. If M_s^* results in worse fit, the model may still be accepted with some non-zero probability. The function in (3) is based on Kirkpatrick and colleagues' work (Kirkpatrick et al., 1983) and is often used in applications of SA, in part due to its similarity to acceptance probabilities under MH (see p. 672).

Acceptance of a suboptimal model may still be useful, especially early in the algorithm, to the extent that it allows SA to avoid being stuck in a local optimum. However, t_s typically decreases across iterations as determined by a cooling schedule so that the probability of accepting a suboptimal model is less likely over time. A conceptual explanation of this behavior is as follows. If after many iterations SA has led M_s to (hopefully) be near the global optimum, a lower value for t_s will provide increasingly smaller acceptance probabilities for suboptimal models, potentially forcing additional accepted models to be closer and closer to the optimum.

Although there is a rich literature on the selection of a starting value and cooling schedule for t_s, in this paper we opted for a simplistic solution as a preliminary test of SA's potential. In particular, we considered starting temperatures of 5, 10, and 25. To provide a concrete example, suppose an increase in BIC of 10 is very undesirable. With $t_s = 5$, $t_s = 10$, and $t_s = 25$ such an increase would yield acceptance of approximately .14, .37 and .67, respectively, meaning that in most cases such a model would be accepted when $t_s = 25$, but rejected when $t_s = 5$ or $t_s = 10$. We chose a straight cooling schedule in which temperature decreases linearly across iterations: $t_s = t_0(S - s)/S$, where t_0 is the starting temperature. Though we note that finer tuning may result in slightly better performance (Stander & Silverman, 1994).

2 Simulations

Simulations were conducted to test the ability of SA to select polynomial order for MP-based item models. The main outcome was item response function recovery, followed by whether SA correctly modeled non-standard items with an MP. A final purpose was to test MP-based models along with SA under conditions that might occur with a planned missing data design.

2.1 Method

Fixed Factors. Simulated datasets included 100 dichotomous items, 5000 respondents, and a standard normal θ. Twenty-five replications per cell of the below data generation design were conducted, with data generation in R (R Core Team, 2015) and models fitted using *rpf* Pritikin (2016) and *OpenMx* Neale et al. (2016).

Data Generation. We manipulated the percentage of items that followed a non-standard model (20, 40, 60, and 80%), with such IRFs generated as the cumula-

tive distribution function (CDF) from a mixture of normal variates, $p_1\mathcal{N}(\mu_1, \sigma_1^2) + p_2\mathcal{N}(\mu_2, \sigma_2^2) + p_3\mathcal{N}(\mu_3, \sigma_3^2)$. To generate variety in IRFs across items and datasets, the following values were randomly generated, $p_1 \sim \text{unif}(.1, .6)$, $p_2 \sim \text{unif}(.1, .3)$, $p_3 = 1 - p_1 - p_2$, $\mu_1 \sim \mathcal{N}(-2.2, .2^2)$, $\mu_2 \sim \mathcal{N}(2.2, .2^2)$, $\mu_3 \sim \mathcal{N}(0, .2^2)$, $\sigma_1 \sim \mathcal{N}(2, .3^2)$, $\sigma_2 \sim \mathcal{N}(.6, .3^2)$, and $\sigma_3 \sim \mathcal{N}(.6, .3^2)$. The remaining items followed a normal CDF (analogous to a normal ogive model) with $\mu \sim \text{unif}(-2.5, 2.5)$ and $\sigma \sim \mathcal{N}(2, .4^2)$. Although the MP-based item model does not strictly follow the exact same shape as the normal CDF items, we still consider them "standard" items for the following investigation since these items should be well approximated by a 2PL or MP with $k = 0$.

We also compared complete data (all respondents completed all items) versus missing data. The missing data condition emulated a planned missing data design with a balanced incomplete block design. The 100 items were split into 5 blocks of 20 items each. Ten test forms with 40 items (i.e., 2 blocks) each were created, corresponding to 60% missing data. We argue that this number of items per test-taker is not atypical of such a design for item calibration, while 60% missing data at this sample size may pose a challenge for MP models.

Fitted Models. To all datasets, we used the logistic function of an MP in (1) as parameterized by Falk and Cai (2016a) and included three models in which k was fixed to the same for all items: $k = 0$, $k = 1$, and $k = 2$. Note that $k = 0$ corresponds to the 2PL model. Following these models, we performed several runs of SA by crossing the following conditions: Energy (AIC vs. BIC), starting temperature (5, 10, and 25), and number of items to perturb (1 vs. 2). One of the three fixed models with the best energy was chosen as the starting model for each SA run. For all MP models with $k > 0$, soft Bayesian priors following Falk and Cai (2016a) were used. One additional model followed the same procedure as SA and started at the best fixed k model according to BIC, but all candidate models were accepted with certainty. We refer to this approach as *semi-random* in what follows, and was included to test whether SA has any advantage over a completely random search in the neighborhood of the best BIC of the fixed models. This and all SA runs included only 300 iterations, and the best model according to AIC or BIC was recorded as the selected model, regardless of whether it was the last accepted model.

2.2 Results

Response Function Recovery. Recovery of response functions was assessed using root integrated mean square error (RIMSE) (Ramsay, 1991), using $Q = 101$ equally spaced quadrature points (X_q) between -5 and 5:

$$\text{RIMSE}_j = \left(\frac{\sum_{q=1}^{Q} ((\hat{P}_j(1|X_q) - P_j(1|X_q))^2 \phi(X_q)}{\sum_{q=1}^{Q} \phi(X_q)} \right)^{1/2} \times 100 \qquad (4)$$

which can be understood as the root of a weighted mean of squared differences between true, $P_j(1|X_q)$, and estimated, $\hat{P}_j(1|X_q)$, response functions, with the population density for θ, $\phi(X_q)$, providing weights. Lower values of RIMSE are better, and the values we report were averaged across all items and replications in each cell of the simulation design.

In general, differences across most tuning options for SA were small for RIMSE, with the number of item perturbations and starting temperature resulting in differences in average RIMSE less than .1 in each cell of the data generation design. For brevity, we report RIMSE results using an initial temperature of 5 and a single item perturbation per iteration of the algorithm. This starting temperature slightly outperformed the other SA conditions.

The best (according to AIC or BIC) out of the fixed (all $k = 0$, $k = 1$, or $k = 2$) models was compared with the 2PL (all $k = 0$), SA (using AIC or BIC), and semi-random models, and RIMSE for these models appears in Table 1. The best performing model is highlighted in bold for each column, and the second best in bold and italics. We highlight several noticeable trends. First, AIC tended to do better than BIC with complete data and a higher percentage of non-standard items. This result tended to hold regardless of whether SA or a fixed k was utilized. For example, with complete data and 80% non-standard items, use of AIC resulted in RIMSE of 1.98 and 1.95, for SA and fixed conditions, respectively, whereas BIC resulted in 2.42 and 2.91. With only 20% non-standard items, BIC performed better than AIC, and this was especially true under missing data where SA using AIC (RIMSE = 1.93) had worse performance than SA with BIC (RIMSE = 1.77) and all other models (RIMSE = 1.81). SA in conjunction with AIC selection was otherwise the best or second best performing model across all other conditions. However, we note that SA with BIC *always* outperformed the 2PL and semi-random conditions. In contrast, SA with

Table 1 Root integrated mean square error (response function recovery)

	Complete data				Missing data			
Model	20%	40%	60%	80%	20%	40%	60%	80%
SA (AIC)	***1.45***	**1.57**	**1.81**	***1.98***	1.93	***2.31***	**2.57**	**2.75**
SA (BIC)	**1.39**	1.70	2.09	2.42	**1.77**	**2.26**	***2.66***	3.02
Fixed (AIC)	1.58	***1.65***	***1.82***	**1.95**	***1.81***	2.34	2.71	***2.93***
Fixed (BIC)	1.50	1.98	2.50	2.91	***1.81***	2.33	2.80	3.24
2PL	1.50	1.98	2.50	3.01	***1.81***	2.33	2.80	3.24
Semi-random	1.49	1.96	2.43	2.81	***1.81***	2.33	2.80	3.23

Note Percentages refer to the number of non-standard true item response models. *SA* Simulated annealing; fixed = best out of all $k = 0$, $k = 1$, $k = 2$, models according to AIC or BIC; 2PL = two-parameter logistic. The best RIMSE value in each column is in bold, the second best is in bold and italics

AIC had poor performance in this single cell of the design versus the 2PL and semi-random model. Finally, SA tended to do better than use of fixed k, though this trend tended to hold within a particular information criterion. For instance, SA with AIC tended to do better than fixed k with AIC selection, and SA with BIC did better than fixed k with BIC selection.

Flagging of Non-standard Items. Although a secondary outcome, we might expect that better fitting models using SA will tend to have non-standard items modeled using $k > 0$. We therefore examined *sensitivity* = # hits/# actual positives = # non-standard items using MP/# non-standard items, and the *false positive rate* = # false positives/# actual negatives = # standard items using MP/# standard items.

We desire high sensitivity, but low false positives—the upper-left region of each panel in Fig. 2. A starting temperature of 5 had a slight advantage over 10, which in turn was better than 25. A better sensitivity/false positive trade-off appears present under complete versus missing data. AIC (versus BIC) resulted in higher sensitivity, but also more false positives. It is difficult to further compare AIC and BIC due to little overlap on each axis. In some cases BIC had near zero false positives, but enough sensitivity to improve IRF recovery. For BIC and a starting temperature of 5, only two cells had false positive rates above .02 (both complete data, 80% non-standard, with .16 and .19). Excluding these two cells, sensitivity for BIC still ranged from .07 to .41. Although not explicitly depicted, a lower percentage of non-standard items tended towards the lower left of these plots, and increasing percentages are connected by straight lines. That is, a higher percentage of non-standard items tended to result in higher sensitivity and higher false positives.

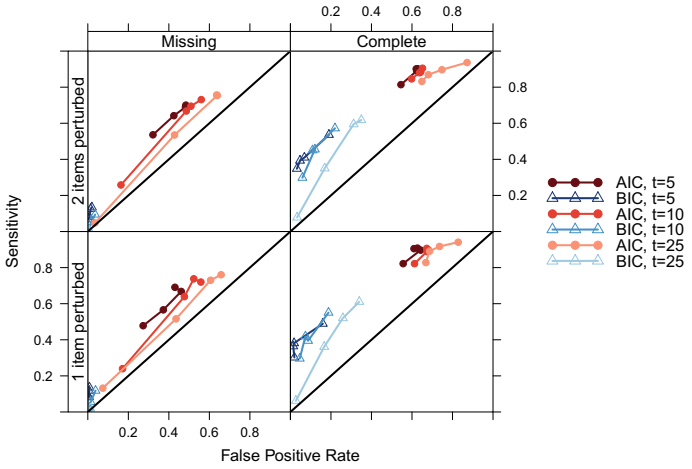

Fig. 2 False positives and sensitivity for final models selected by simulated annealing. *Note* "*t*" indicates starting temperature

3 Discussion and Conclusion

We conclude that SA has potential to aid in selecting polynomial order for MP-based item models in that SA tended to improve IRF recovery under most conditions. This result is promising given our initial attempt at SA implementation. For instance, tuning of the cooling schedule may further improve performance. In retrospect, a starting temperature of 25 may allow initial acceptance of many poorly fitting models, and a lower starting temperature is preferable. The number of iterations could also be increased, yet a computational advantage is still apparent over a step-wise approach: 300 fitted models would have only allowed change in polynomial order for 3 items on a test with $n = 100$.

There were some trade-offs in the choice of AIC versus BIC. AIC tended to have greater gains in IRF recovery, except under missing data and when few items followed a non-standard model. As AIC had greater sensitivity in modeling non-standard items with an MP, it also tended to result in some over-fitting. Given the great contrast in sensitivity and false positive rates, we suppose that the psychometrician's preference for a conservative (BIC) versus liberal (AIC) flagging of non-standard items may guide which to use. Of course, other optimality criterion or use of other item fit statistics may be used in future research. In addition, test length, sample size, and the amount of missing data may also be important to consider and could be further examined.

A similar computational problem may hold for other flexible parametric modeling techniques (Lee & Bolt, 2018): Should we use a standard item model or a different modeling approach? To the extent that the test is very long, this same problem may occur if one is trying to decide between several different models for *each* test item. Of course, substantive theory should be used to guide modeling choices where possible. However, in the absence of good theory, an automated approach such as that we have provided here may be a reasonable start to help improve fit while identifying which items require closer examination, especially for a long test or large item bank. MP-based models do not directly inform about the source of item misfit. Further follow-up analyses with alternative models and/or content analysis of particular items may provide insight into whether an MP or other modeling approach is appropriate. That is, there is both room for MP-based item models to complement other modeling approaches, and also for such combinatorial optimization algorithms to be used in selecting whether to use any of these other modeling approaches.

Acknowledgements We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number RGPIN-2018-05357]. Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), [numéro de référence RGPIN-2018-05357].

References

Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability* (pp. 395–479). Reading, MA: Addison-Wesley.

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

Černý, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, *45*, 41–51.

Drezner, Z., & Marcoulides, G. A. (1999). Using simulated annealing for selection in multiple regression. *Multiple Linear Regression Viewpoints*, *25*, 1–4.

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

Edwards, M. C., Flora, D. B., & Thissen, D. (2012). Multistage computerized adaptive testing with uniform item exposure. *Applied Measurement in Education*, *25*, 118–114.

Falk, C. F. (2018). *A monotonic polynomial graded response model*. Presentation at the International Test Commission Conference, Montreal, Canada.

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. (2016). Exploring alternate latent trait metrics with the filtered monotonic polynomial IRT model. Ph.D. thesis, Department of Psychology, University of Minnesota.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–680.

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

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

Molenaar, D. (2015). Heteroscedastic latent trait models for dichotomous data. *Psychometrika*, *80*(3), 625–644.

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kickpatrick, R. M., et al. (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, r package version 0.53.

R Core Team. (2015). R: A language and environment for statistical computing. http://www.R-project.org, ISBN 3-900051-07-0.

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

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

Stander, J., & Silverman, B. W. (1994). Temperature schedules for simulated annealing. *Statistics and Computing*, *4*, 21–32.

Woods, C. M. (2007). Empirical histograms in item response theory with ordinal data. *Educational and Psychological Measurement*, *67*, 73–87.
Xu, X., & Douglas, J. A. (2006). Computerized adaptive testing under nonparametric IRT models. *Psychometrika*, *71*(1), 121–137.