PSYCHOMETRIKA—VOL. 85, NO. 2, 301–321
JUNE 2020
https://doi.org/10.1007/s11336-020-09703-8

A SHADOW-TEST APPROACH TO ADAPTIVE ITEM CALIBRATION

WIM J. VAN DER LINDEN
UNIVERSITY OF TWENTE
BINGNAN JIANG
ACT, INC.

A shadow-test approach to the calibration of field-test items embedded in adaptive testing is presented. The objective function used in the shadow-test model selects both the operational and field-test items adaptively using a Bayesian version of the criterion of D_S-optimality. The constraint set for the model can be used to hide the field-test items completely in the content of the test as well as to deal with such practical issues as random control of their exposure rates. The approach runs on efficient implementations of the Gibbs sampler for the real-time updating of the ability and field-test parameters. Optimal settings for the proposed algorithms were found and used to demonstrate item calibration with smaller than traditional sample sizes in runtimes fully comparable with conventional adaptive testing.

Key words: adaptive testing, Bayesian D_S-optimality, Gibbs sampling, item calibration, item response models, MCMC algorithm, shadow-test approach.

1. Introduction

The current focus of adaptive testing programs typically is on the selection of items from an operational pool with optimization of the estimates of the examinees' ability parameters as statistical objective. As for the item pool, it is not uncommon for it to be calibrated with response data collected in separate field-test studies with the items organized as sets of linked test forms administered to volunteering examinees. This type of field-test studies requires considerable resources. It may also run into such problems as less representative data due to lack of the operational testing conditions, danger of early security breaches, detection of misfitting items only after all data have been collected, and drifting of the item parameters from the scales for the current item pool due to accumulated linking errors.

An alternative approach is to embed a few field-test items directly in the operational adaptive tests for each of the examinees. A unique advantage possible for this approach is adaptive selection of the items capitalizing on the real-time updates of the examinees' ability parameters during testing. Just as adaptive selection of the operational items leads to a considerable reduction of the test length necessary to score the examinees, similar selection of the field-test items can be expected to lead to substantial reduction of the sample size necessary to calibrate new items, an advantage already demonstrated in a few recent item calibration studies (Ren et al. 2017; van der Linden and Ren 2015). Other possible advantages of embedded field-testing of new items are more practical and include saving the time and expenses necessary to run separate calibration studies, collection of responses from motivated examinees answering the items under operational conditions, possibility of real-time monitoring of item fit and early withdrawal of items that appear

Correspondence should be made to Wim J. van der Linden, University of Twente, Enschede, The Netherlands. Email: wjvdlinden@outlook.com

© 2020 The Author(s)

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

to be malfunctioning, and estimation of all item parameters directly on the scales in use for the item pool.

These advantages are only possible when we have algorithms that manage the processes of parameter updating and optimal item selection both for the operational and field-test items in real time. Also, some additional practical issues need to be resolved. For example, the identity of the field-test items should be hidden from the examinees. If they would be able to find out which items do and do not add to their scores, they may take the latter less seriously and even decide to answer them just randomly to save time for the items that do count. Also, as field-test items are now exposed to their future population of examinees prior to operational testing, it may be necessary to prevent their possible compromise due to this extra exposure. An obvious option is to subject them to the same type of shadow-test-based exposure control as for the operational items in the pool (van der Linden and Choi 2019) but with adjusted values for their target exposure rates.

The goal of the research in this report was to design and optimize the algorithms required for real-time adaptive testing with embedded item calibration. The approach is derived from a more general description of adaptive testing programs as multiple, statistical and non-statistical processes with real-time updating of all intentional statistical parameters in van der Linden (2018). The tools available to run these processes are sequential Bayesian parameter updating, application of statistical optimal design theory to plan each next data collection step, and use of the shadow-test approach to manage all processes simultaneously and keep each of the non-statistical parameters of the program within their required bounds.

The basic setup of the testing program considered in the current research was an adaptive test with a few field-test items administered to each of the examinees in positions randomly selected near the end of the test (for example, three of the last five). The choice of setup was motivated by the necessity to compromise between the ideal of selecting the field-test items capitalizing on the statistical information offered by the posterior distributions of the examinees' ability parameters and the requirement to hide their identity to have them taken equally seriously as the operational items in the test. As explained below, the identity of the field-test items can further be hidden by embedding them fully in the test content. The compromise between the wish to have maximum statistical information and to hide the identity of the field-test items may need to be extended to account for the effects of possible fatigue or speededness of the test as well, an issue reserved for future research.

The following sections introduce the response model used as an example in the research, explain each of the three tools referred to above in more detail, and present results from studies conducted to optimize the settings of the algorithms and demonstrate their application to a real-world adaptive testing program. All results point at item calibration with smaller sample sizes than for separate calibration studies obtained in runtimes fully comparable with those for current maximum-information adaptive testing without any field testing of new items.

2. Response Model

Let $i = 1, \ldots, I$ denote the operational items and $f = 1, \ldots, F$ the field-test items in the pool. Both categories of items are scored by response variables $U_i \in \{0, 1\}$ and $U_f \in \{0, 1\}$, respectively. As an example, the operational items are assumed to be calibrated and their fit checked under the three-parameter logistic (3PL) model, which explains the probability of a correct response on each of the items as

$$\Pr\{U_i = 1|\theta, a_i, b_i, c_i\} \equiv p(\theta; a_i, b_i, c_i) \equiv c_i + (1 - c_i)\frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \qquad (1)$$

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

where $b_i \in \mathbb{R}$ and $a_i \in \mathbb{R}^+$ can be interpreted as parameters for the difficulty and discriminating power of item i, respectively, and $c_i \in (0, 1]$ as the probability of a correct response to the item resulting from purely random guessing. The model allows us to write the probability function of the response distribution for each examinee–item combination as

$$f(u_i|\theta, \boldsymbol{\xi}_i) \equiv p(\theta, \boldsymbol{\xi}_i)^{u_i}[1 - p(\theta, \boldsymbol{\xi}_i)]^{1-u_i}, \tag{2}$$

where $\boldsymbol{\xi}_i \equiv (a_i, b_i, c_i)$. The field-test items are calibrated under the same model. It is prudent to monitor their fit to the response model permanently during the calibration process to prevent the examinees from having to spend unnecessary time on malfunctioning items. The question of how to adjust goodness-of-fit statistics for application in real-time monitoring is currently investigated.

One of the novel aspects of the approach outlined below is that, rather than storing fixed point estimates for all parameter in the adaptive testing system, each of them is permanently represented by a short vector of draws from its most recent posterior distribution. Due to this decision, all calculations during operational testing simplify with results that automatically account for the uncertainty about any of the parameters. The best way to collect these draws during initial item pool calibration is through Bayesian estimation with Markov chain Monte Carlo (MCMC) sampling of the posterior distributions of the parameters. Alternatively, for an existing item pool, we could use an asymptotic argument sampling normal distributions with the point estimates and standard errors of the item parameters as mean and standard derivation. This is a one-time requirement only, though. Once the process of embedded item calibration begins, the system automatically generates new items ready for operational use along with appropriate vectors of posterior draws for each of them.

3. Sequential Bayesian Parameter Updating

During testing, two distinct processes of parameter updating are to be managed, one for the ability parameters of the examinees and the other for the parameters of the field-test items. These parameters are the intentional parameters of the two processes. In addition, both processes have nuisance parameters. At the update of an ability parameter after a response to an operational item, the parameters of the item are nuisance parameters. But when the parameters of a field-test item are updated, the ability parameter of the examinee that produced the response changes its status from intentional to nuisance parameter. The status of parameters as nuisance parameters does not mean that we can ignore them; just as for the intentional parameters, we need to fully account for their impact on the responses that are collected. Neither is it correct to substitute point estimates for them; doing so would lead to loss of optimality of the criteria for the selection of the next operational and field-test items, especially early on in the two processes when the uncertainty about the ability and field-test parameters is maximal (for an example of the initial uncertainty about latter, see the first lines in Figs. 2, 3, 4).

The appropriate way to update intentional parameters in a continuous processes of response collection is through sequential use of Bayes theorem integrating out each of nuisance parameters. In the current context, after the kth operational item in the test and response vector $\boldsymbol{u}_k = (u_1, \ldots, u_k)$, the theorem is used to update the marginal posterior distribution of the ability parameter from $f(\theta|\boldsymbol{u}_{k-1})$ to $f(\theta|\boldsymbol{u}_k)$. The prior distributions for the operational parameters $\boldsymbol{\xi}_k$ during the update are their posterior distributions obtained during item pool calibration. As these distributions do not depend on any data collected during testing, we have prior independence of $f(\theta \mid \boldsymbol{u}_{k-1})$ and $f(\boldsymbol{\xi}_k)$. Besides, the standard assumption of local independence implies $f(u_k \mid \theta, \boldsymbol{u}_{k-1}) = f(u_k \mid \theta)$. Hence, the application of Bayes theorem, in its regular form equal to

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

$$f(\theta \mid u_k) = \frac{\int f(u_k \mid \theta, \boldsymbol{\xi}_k) f(\theta, \boldsymbol{\xi}_k \mid \boldsymbol{u}_{k-1}) \mathrm{d}\boldsymbol{\xi}_k}{\int\int f(u_k \mid \theta, \boldsymbol{\xi}_k) f(\theta, \boldsymbol{\xi}_k \mid \boldsymbol{u}_{k-1}) \mathrm{d}\theta \mathrm{d}\boldsymbol{\xi}_k},$$

simplifies to

$$f(\theta|\boldsymbol{u}_k) = \frac{f(u_k|\theta)}{f(u_k)} f(\theta|\boldsymbol{u}_{k-1}), \tag{3}$$

where

$$\frac{f(u_k|\theta)}{f(u_k)} = \frac{\int f(u_k|\theta, \boldsymbol{\xi}_k) f(\boldsymbol{\xi}_k) \mathrm{d}\boldsymbol{\xi}_k}{\int\int f(u_k|\theta, \boldsymbol{\xi}_k) f(\boldsymbol{\xi}_k) f(\theta|\boldsymbol{u}_{k-1}) \mathrm{d}\theta \mathrm{d}\boldsymbol{\xi}_{k-1}}. \tag{4}$$

and $f(u_k|\theta, \boldsymbol{\xi}_k)$ is the examinee's response probabilities in (2) for the kth item in the test. Observe how nicely (3) factors the posterior density of θ into the product of one factor depending on the new response and a second summarizing the information about it contained in all earlier responses.

Likewise, the update of the joint posterior distribution of field-test parameter $\boldsymbol{\xi}_f$ upon the lth administration of the item is

$$f(\boldsymbol{\xi}_f|\boldsymbol{u}_l) = \frac{f(u_l|\boldsymbol{\xi}_f)}{f(u_l)} f(\boldsymbol{\xi}_f|\boldsymbol{u}_{l-1}), \tag{5}$$

where

$$\frac{f(u_l|\boldsymbol{\xi}_f)}{f(u_l)} = \frac{\int f(u_l|\theta_l, \boldsymbol{\xi}_f) f(\theta_l|\boldsymbol{u}_{k'}) \mathrm{d}\theta_l}{\int\int f(u_l|\theta_l, \boldsymbol{\xi}_f) f(\boldsymbol{\xi}_f|\boldsymbol{u}_{l-1}) f(\theta_l|\boldsymbol{u}_{k'}) \mathrm{d}\theta_l \mathrm{d}\boldsymbol{\xi}_f}, \tag{6}$$

k' is the number of operational items already answered by the current examinee, and $f(u_l|\theta_l, \boldsymbol{\xi}_f)$ now is the response probability of the examinee in (2) for field-test item f.

For both processes of parameter updating we thus have a simple recurrence relation between the new and previous posterior density, with the latter serving as the new prior. In spite of their apparent simplicity, because of the presence of multiple integrals in (4) and (6), both relations may seem computational demanding. However, as is well known now, use of Monte Carlo sampling from the joint posterior distribution of all parameters avoids having to calculate any of these integrals (e.g., Gelman et al. 2014, Part 3; Gilks et al. 1996). Particularly attractive is the use of the Gibbs sampler which, without any further assumptions, allows for an extremely efficient implementation in the current context of adaptive testing.

Recall that the parameters of the operational items are represented in the system by vectors of posterior draws collected during their calibration. The Gibbs sampler capitalizes on their presence in the first of the following two updates:

Update of Ability Parameter θ

1. The sampler iterates between resampling of the vectors of draws for the parameters $\boldsymbol{\xi}_k$ of the last operational item administered to the examinee and a Metropolis–Hastings (MH) step to sample from the new distribution of the θ parameter.
2. When the sampler is stopped, the current vector of draws for θ in the system is overwritten with a selection of new draws collected from the stationary part of the Markov chain.

Update of Field-Test Parameters $\boldsymbol{\xi}_f$

1. The sampler now iterates between resampling of the last vectors of draws for the θ parameter of the current examinee and an MH step to sample the new distribution of field-test parameter $\boldsymbol{\xi}_f$.
2. When the sampler is stopped, the current vector of draws for $\boldsymbol{\xi}_f$ in the system is overwritten with a selection of new draws collected from the stationary part of the Markov chain.

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

The sequential nature of both updates permits a natural choice of proposal and prior distributions for their MH steps. For instance, for the update of the θ parameter, let $(\theta_{k-1}^{(1)}, \ldots, \theta_{k-1}^{(S)})$ be the vector of S posterior draws currently present in the system, which has mean and variance

$$\mu_{k-1} = S^{-1} \sum_{s=1}^{S} \theta_{k-1}^{(s)} \tag{7}$$

and

$$\sigma_{k-1}^2 = S^{-1} \sum_{s=1}^{S} \left(\theta_{k-1}^{(s)} - \mu_{k-1} \right)^2, \tag{8}$$

respectively. As the MH steps require specification of the prior density of each of the values of θ drawn from the posterior distribution, which is known to converge to normality anyhow (Chang and Ying 2009), an obvious choice is

$$f_{k-1}(\theta \mid \boldsymbol{u}_{k-1}) \equiv N(\mu_{k-1}, \sigma_{k-1}^2). \tag{9}$$

Likewise, for the proposal distribution at iteration steps $r = 1, \ldots, R$, an effective choice is

$$q_k(\theta \mid \theta^{(r-1)}) \equiv N(\theta^{(r-1)}, \sigma_{k-1}^2). \tag{10}$$

Because of the symmetry of (10), upon the rth draw for $\boldsymbol{\xi}_k$, the draw for the θ parameter simplifies to

(1) draw candidate value $\theta^{(c)}$ for $\theta^{(r)}$ from the proposal distribution in (10);
(2) accept $\theta^{(r)} = \theta^{(c)}$ with probability

$$\min \left\{ \frac{N(\theta^{(c)} \mid \mu_{k-1}, \sigma_{k-1}^2) f(u_k; \theta^{(c)}, \boldsymbol{\xi}_k^{(r)})}{N(\theta^{(r-1)} \mid \mu_{k-1}, \sigma_{k-1}^2) f(u_k; \theta^{(r-1)}, \boldsymbol{\xi}_k^{r-1})}, 1 \right\}; \tag{11}$$

otherwise $\theta^{(r)} \equiv \theta^{(r-1)}$.

As the Markov chain is already on target right from its start for each of the nuisance parameters, convergence for the single intentional parameter is extremely fast. Also, the only quantity that needs to be calculated to evaluate (11) is the product of a normal density and the response probability for the last item in its numerator (the denominator was already calculated in the previous step). Finally, it is not necessary to tune the proposal distribution, typically a time-consuming task. The distribution automatically follows the posterior distribution with a somewhat greater variance, known to be an efficient choice for low-dimensional parameters (Gelman et al. 2014, sect. 12.2). For further details, such as the optimal choice of burn-in length, autocorrelation estimates, and control of Monte Carlo error, the reader should consult the optimization study of the sampler for application in regular adaptive testing without any item calibration in van der Linden and Ren (2020).

The sampler for the updates of each of the field-test parameters added in this research follows the same setup as in (7)–(11). The only difference is a temporary change of their scale to support the normality of (9) and (10). Prior to the update, the parameters are transformed to

$$b_f^* \equiv b_f$$

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

$$a_f^* \equiv \ln(a_f)$$
$$c_f^* \equiv \text{logit}(c_f)$$

with back transformation to the original scale immediately after the update. The authors are aware of the fact that, for the multi-parameter case, direct sampling from the joint posterior distribution is generally more efficient. But, given the already extremely fast convergence, the gains for the current application have been found to be too small to justify the efforts.

4. Optimal Design Criteria

The ultimate goal of the two processes of adaptive testing and item calibration is maximum information about their intentional parameters. To realize this, sequential updating of these parameters is not sufficient. The necessary additional step is selection of each next item to have the best possible match between the intentional and nuisance parameters for the pertinent process.

Criteria for such matches can be derived from statistical optimal design theory. Among its many criteria, the more favorable belong to the class of D_s-optimality; that is, the choice of a minimum determinant of the covariance matrix of the estimators of the intentional parameters given the nuisance parameters. For a general introduction to this class of criteria from a frequentist perspective, refer to Holling and Schwabe (2018), while Berger (2018) should be consulted for a general introduction to optimal item-calibration design. Alternative optimal design criteria for use in adaptive item calibration have been studied by Ren et al. (2017) and van der Linden and Ren (2015) but these authors also found the criterion of D_s-optimality to serve its purpose generally best. Minimization of the determinant of the covariance matrix for maximum-likelihood estimators is (asymptotically) equivalent to maximization of their Fisher information matrix. As the items are selected in a continuous process of data collection, the proposed version of the D_s-criterion is selection of items with the maximum marginal profit for either of these determinants.

For θ as intentional parameter, the information matrix reduces to a scalar (equal to its "determinant") widely used as item-selection criterion in conventional adaptive testing. For the 3PL model, it is easily obtained as

$$I(\theta; \boldsymbol{\xi}_i) = a_i^2 \frac{1 - p(\theta; \boldsymbol{\xi}_i)}{p(\theta; \boldsymbol{\xi}_i)} \left(\frac{p(\theta; \boldsymbol{\xi}_i) - c_i}{1 - c_i} \right)^2. \tag{12}$$

Because of local independence, (12) is additive in the items and consequently all contributions by the earlier items can be ignored when selecting the next. However, as both θ and $\boldsymbol{\xi}_i$ are known only through their posterior distributions, the proper way of using (12) is by first taking its expectation across these distributions. The proposed Gibbs sampler simplifies the calculation of the expectation considerably. In addition to the vector $(\theta^{(1)}, \ldots, \theta^{(S)})$ of draws from last update of θ, let $(\boldsymbol{\xi}_i^{(1)}, \ldots, \boldsymbol{\xi}_i^{(S)})$ be the vector for the parameters of operational item i permanently present in the system. Using these draws, the posterior expected information for i as candidate item can be calculated as

$$\int \int I(\theta; \boldsymbol{\xi}_i) f(\theta \mid u_k) f(\boldsymbol{\xi}_i) \mathrm{d}\theta \mathrm{d}\boldsymbol{\xi}_i \approx S^{-1} \sum_{s=1}^{S} I(\theta^{(s)}, \boldsymbol{\xi}_i^{(s)}). \tag{13}$$

The length of both vectors of draws is taken to be equal here for notational convenience only. For vectors of different lengths, an efficient choice is to recycle the shorter against the longer. The

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

length of the vectors directly controls the Monte Carlo error in (13), an issue further addresses below.

As for the selection of the field-test items, the Fisher information matrix for field-test item f is equal to

$$I(\boldsymbol{\xi}_f;\theta) = \left(-\frac{\partial^2}{\partial\xi_j\partial\xi_{j'}}\ln(f(u;\theta,\boldsymbol{\xi}_f)\right), \quad j, j' \in \{1,2,3\}. \tag{14}$$

Using

$$\lambda_f \equiv \frac{(1-p_f)}{p_f(1-c_f)^2}$$

with $p_f \equiv f(u;\theta,\boldsymbol{\xi}_f)$, the elements of (14) are easily obtained from

$$\frac{\partial^2}{\partial a_f^2}\ln(f(u;\theta,\boldsymbol{\xi}_f)) = -(p_f-c_f)^2(\theta-b_f)^2\lambda_f, \tag{15}$$

$$\frac{\partial^2}{\partial b_f^2}\ln(f(u;\theta,\boldsymbol{\xi}_f)) = -a_f^2(p_f-c_f)^2\lambda_f, \tag{16}$$

$$\frac{\partial^2}{\partial c_f^2}\ln(f(u;\theta,\boldsymbol{\xi}_f)) = -\lambda_f, \tag{17}$$

$$\frac{\partial^2}{\partial a_f b_f}\ln(f(u;\theta,\boldsymbol{\xi}_f)) = a_f(p_f-c_f)^2(\theta-b_f)\lambda_f, \tag{18}$$

$$\frac{\partial^2}{\partial a_f c_f}\ln(f(u;\theta,\boldsymbol{\xi}_f)) = -(p_f-c_f)(\theta-b_f)\lambda_f, \tag{19}$$

$$\frac{\partial^2}{\partial b_f\partial c_f}\ln(f(u;\theta,\boldsymbol{\xi}_f)) = a_f(p_f-c_f)\lambda_f. \tag{20}$$

Though the matrix is still additive in the items, its determinant is not. As a result, the expected contribution by the current examinee to each of the candidate items is no longer independent of its history. This fact seems to complicate evaluation of the D_s-criterion for field-test item selection: Using it for the information matrix would require calculation of the posterior expectation of the determinant of the sum of these matrices for each candidate item across the ability parameters of all examinees who have already seen it minus the determinant for the current examinee. On the other hand, direct use of the posterior covariance matrix seems computationally challenging too. It would require simulation of the adaptive test one item ahead for each of the candidate items, running the Gibbs sampler, calculating the determinant of the covariance matrix from the new posterior samples for the parameters of the items, and subtracting the determinant calculated from their current samples. However, the equivalence of the two approaches allows us to capitalize on the data about their key quantities already available in the system.

Let $\text{Cov}(\boldsymbol{\xi}_f)$ denote the covariance matrix calculated from the last update of the posterior draws for field-test parameters $\boldsymbol{\xi}_f$ and $I(\boldsymbol{\xi}_f;\theta)$ the information matrix for the current examinee and candidate item f as defined by (14)–(20). It is possible to write the criterion as

$$D_s(\boldsymbol{\xi}_f) = \det\left(\text{Cov}^{-1}(\boldsymbol{\xi}_f) + I(\boldsymbol{\xi}_f;\theta)\right) - \det\left(\text{Cov}^{-1}(\boldsymbol{\xi}_f)\right), \tag{21}$$

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

while its posterior expected value can be calculated using

$$D_{\mathrm{s}}(\boldsymbol{\xi}_{\boldsymbol{f}}) \approx S^{-1} \sum_{s=1}^{S} \left[\det\left(\mathrm{Cov}^{-1}(\boldsymbol{\xi}_f) + I(\boldsymbol{\xi}_f^{(s)}; \theta^{(s)}) \right) - \det\left(\mathrm{Cov}^{-1}(\boldsymbol{\xi}_f) \right) \right]. \tag{22}$$

4.1. Shadow-Test Approach

In the shadow-test approach to adaptive testing (van der Linden and Reese 1998; van der Linden 2005, chap. 9), a full-length shadow test is re-assembled prior to the selection of each item. During each re-assembly, the test must satisfy the following requirements: (i) its objective function guarantees maximum information about the intentional parameters at their current update; (ii) the constraint set covers all test specifications in force for the program; and (iii) all items already administered to the examinee are included in the test. The next item administered to the examinee is the most informative free item in the current shadow test. The examinee only sees these items; all other items remain unknown (hence the name of "shadow test"). A distinct advantage of the approach is optimality of item selection given the constraint set along with guaranteed satisfaction of the latter. A flexible way of implementing the approach is through mixed integer programming (MIP) modeling of the objective function and constraint set with submission of the optimization model to a MIP solver prior to the selection of each item. The solver then returns the IDs of the optimal selection of items, from which the testing algorithm picks the best item for administration. Powerful MIP solvers, fully up to their use in real-time shadow-test assembly, are available as open-source or commercial software programs; information about the runtimes in the empirical studies for the solver used in this research is given below.

Though the approach has only been used for conventional adaptive testing so far, it is actually much more powerful. The objective function and constraint set in the shadow test model can be chosen to manage almost every combination of processes involved in adaptive testing. To do so, we just need to consult the rich body of knowledge and experience present in the field of MIP modeling, especially in its applications to problems of multi-objective decision-making (e.g., Chen et al. 2010; Williams 2013). The remainder of this section illustrates the power of the approach for the simultaneous processes of adaptive testing and item calibration.

Each shadow test is assumed to have n_{o} operational and n_f field-test items. Binary variables $x_i = 0, 1$ and $x_f = 0, 1$ are used to select the two categories of items from the pool. In addition, we use S_{k-1} to denote the set of indices of the $k-1$ operational and field-test items already administered to the examinee when assembling the kth shadow test. Examples of constraints available to deal with possible specifications of the required distributions of quantitative and categorical attributes of the items in the adaptive test are given. Quantitative attributes are attributes with numerical values q that represent such features as their response model parameters, expected response times, exposure rates, etc. The distributions of these attributes are required to be constrained by upper and/or lower bounds b_q. Categorical attributes partition the item pool into subsets V_c with common features $c = 1, \ldots, C$. The number of items selected from these subsets are to be constrained by upper and/or lower bounds n_c. A prime example of a categorical attribute is a (possibly multi-level) classification of the content of the items in the pool; other examples are item type, format, answer keys, etc. Our last example of a possible constraint is a set of items V_e that have to be excluded from the test. One possible use of such sets is item-exposure control through random exclusion of items from the shadow tests for the examinees with probabilities controlled by the empirical exposure rates of the items in the pool.

All these bounds and sets are instances of what we have referred to earlier as non-statistical parameters of an adaptive testing program that need to be maintained during testing. From the point of view of test validity, they are more important than its statistical parameters. An example

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

of a shadow-test model for the optimal selection of the kth item in the test that does maintain the desired values of these parameters is

$$\text{maximize} \sum_{i=1}^{I} I_i(\theta)x_i + \sum_{f=1}^{F} D_{\text{s}}(\boldsymbol{\xi}_f)x_f \quad \text{(objective function)} \tag{23}$$

subject to

$$\sum_{i=1}^{I} x_i = n_{\text{o}}, \quad \text{(\# of operational items)} \tag{24}$$

$$\sum_{f=1}^{F} x_f = n_f, \quad \text{(\# of field-test items)} \tag{25}$$

$$\sum_{i,f \in S_{k-1}} (x_i + x_f) = k - 1, \quad \text{(items already administered)} \tag{26}$$

$$\sum_{i=1}^{I} q_i x_i \lessgtr b_{\text{o}q}, \quad \text{(quantitative constraint)} \tag{27}$$

$$\sum_{f=1}^{F} q_f x_f \lessgtr b_{fq}, \quad \text{(quantitative constraint)} \tag{28}$$

$$\sum_{i \in V_c} x_i \lessgtr n_{\text{o}c}, \; c = 1, \ldots, C, \quad \text{(categorical constraints)} \tag{29}$$

$$\sum_{f \in V_c} x_f \lessgtr n_{fc}, \; c = 1, \ldots, C, \quad \text{(categorical constraints)} \tag{30}$$

$$\sum_{i,f \in V_e} (x_i + x_f) = 0, \quad \text{(exclusion constraint)} \tag{31}$$

$$x_i = 0, 1, \; i = 1, \ldots, I, \quad \text{(binary decision variables)} \tag{32}$$

$$x_f = 0, 1, \; f = 1, \ldots, F. \quad \text{(binary decision variables)} \tag{33}$$

The objective function simultaneously maximizes the posterior expected information in the shadow test about the examinee's ability parameter and the field-test parameters in the pool as calculated by the two criteria in (13) and (22). The first two constraints control the numbers of items to be administered. The constraint in (26) effectively sets the decision variables of the $k - 1$ operational and field-test items already answered by the examinee equal to one. The next two types of constraints are to maintain bounds on quantitative item attributes. Typically, multiple versions of these constraints are necessary both with upper and/or lower bounds. The same holds for the constraints on the categorical item attributes in (29) and (30). Obviously, the zero bound in (31) prevents the items in set V_e in the pool from being administered to the examinee, for instance, because they contain clues to each other ("enemy items"). The last two sets of constraints are necessary to define the decision variables as binary.

The model only serves as an example and does not exhaust all possibilities. A major missing type of constraint are those for logical item selection. These constraints are necessary, for instance, to select set-based items along with their stimuli or prevent the selection of items with an enemy relationship. For these and other modeling options, refer to van der Linden (2005). Also, it is

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

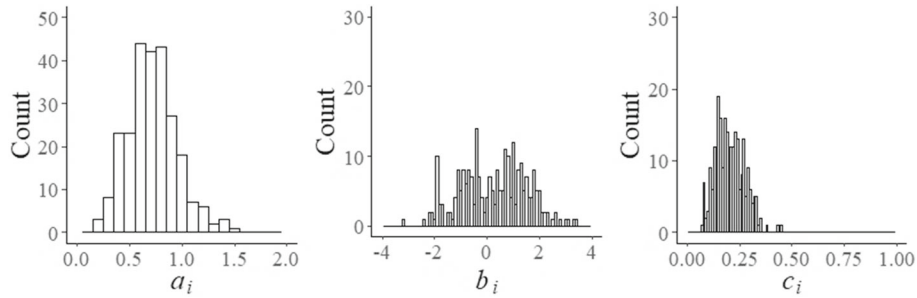


FIGURE 1.
Distributions of parameters of operational items in the pool

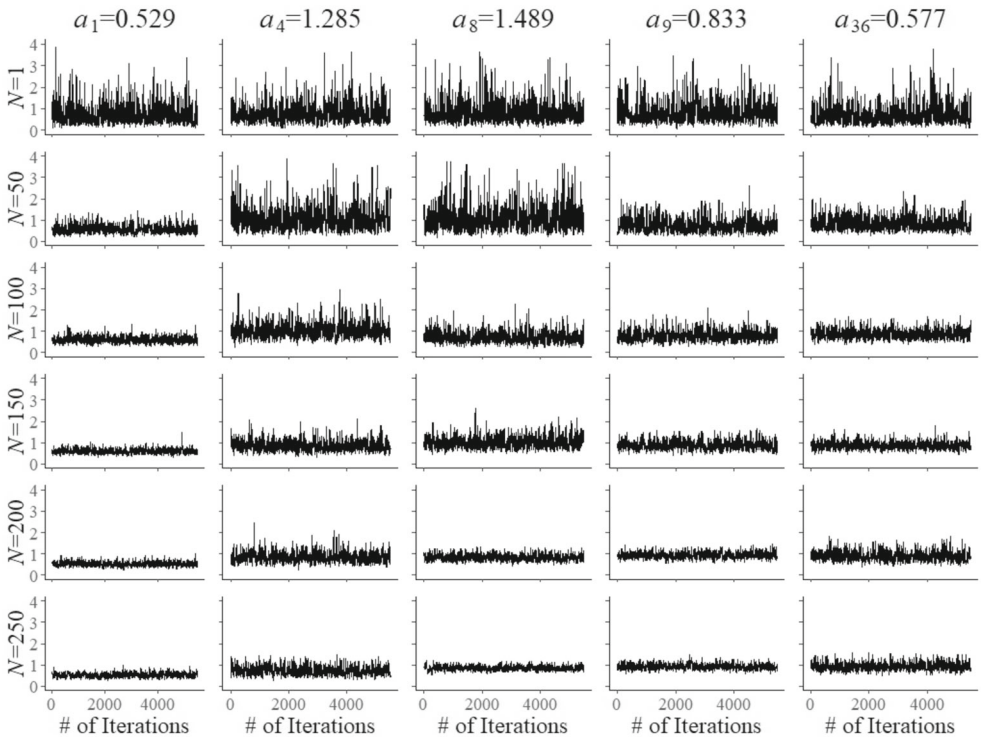


FIGURE 2.
Trace plots of Markov chains for field test parameters a_f of five arbitrary items after $N = 1, 50, 10, 50, 150, 200$, and 250 responses

important to note that the only required update of this model prior to moving to the next item is addition of the index of the item just administered to set S_{k-1} in the left-hand side of (26) along with an increase of item count k in its right-hand side by one.

A key feature of the model in (23)–(33) is complete separation of the optimization of the selection of the operational and field-test items. Though the two problems are controlled by a single model for the shadow test, both have their own objective of maximum information and are subjected to their own bounds. The only difference between the two types of items emerges when an item is selected for administration. When the system arrives at a position in the test reserved

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

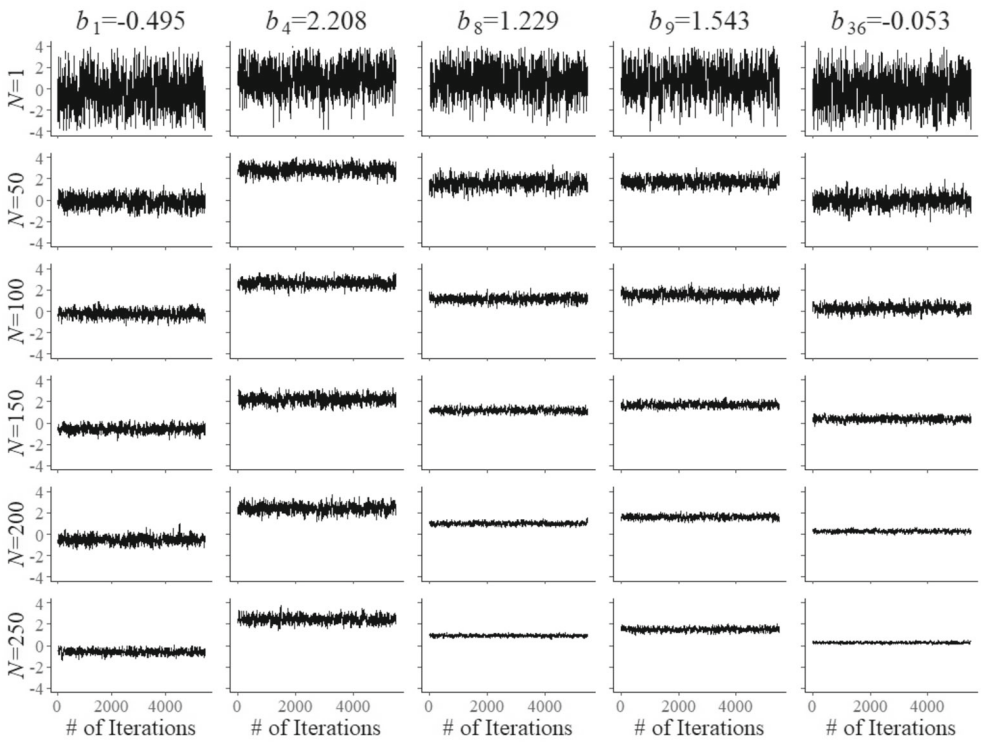


FIGURE 3.
Trace plots of Markov chains for field test parameters b_f of five arbitrary items after $N = 1$, 50, 10, 50, 150, 200, and 250 responses

for item calibration, it is the currently best field-test item in the shadow test that is administered; otherwise, it is the best operational item. It is possible to combine pairs of constraints as in (27)–(30) replacing them with single versions running across both types of items in the pool. The effect is less specific control of the composition of the shadow test, which may be a disadvantage if there exists a shortage of items with certain combinations of attributes and their calibration has to be accelerated. Also, as operational and field-test items then become exchangeable with respect to the attribute in their common constraint, the system may select an item with the attribute from one of these two categories whereas it may have picked one from the other in case of separate constraints (subject to all other constraints, of course).

5. Simulation Study

The goal of the study was to explore optimal settings for the proposed shadow-test approach to embedded item calibration and demonstrate its feasibility for real-world application. More specifically, our goal was to address such issues as the best burn-in length of the Markov chain for the updates of the field-test parameters, estimation of its autocorrelation, selection of the sample size of the post burn-in draws for operational use, as well as the recording of the runtimes for our best combination of settings. In order to demonstrate practical feasibility, the combination was evaluated both for its speed of item calibration and the statistical qualities of the item parameter estimates.

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

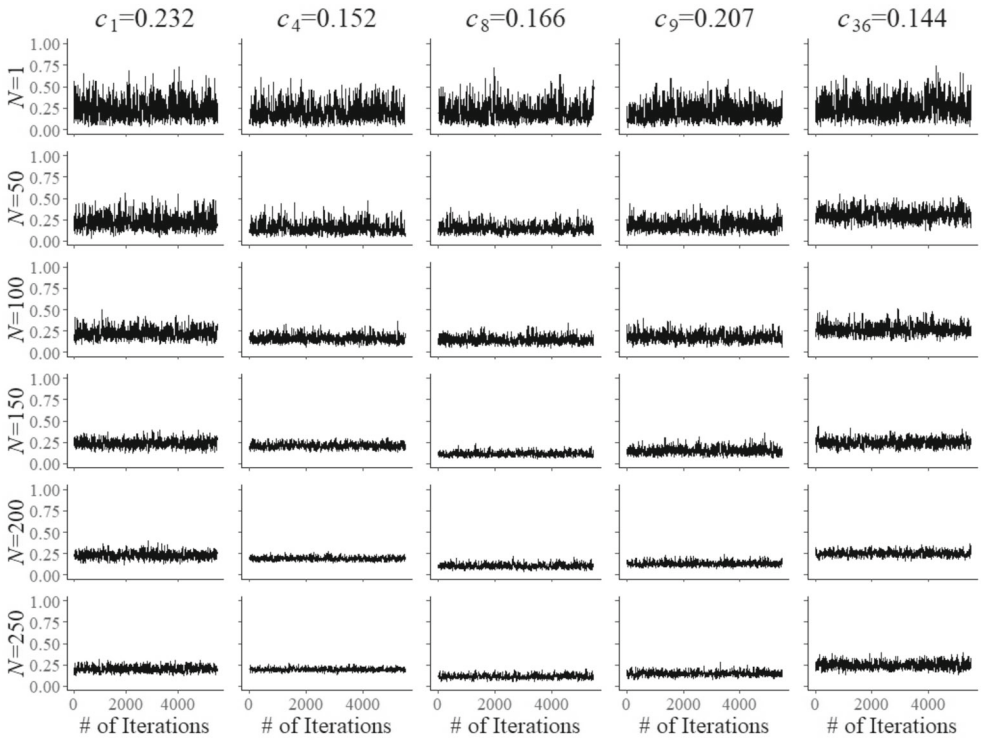


FIGURE 4.
Trace plots of Markov chains for field test parameters c_f of five arbitrary items after $N = 1$, 50, 10, 50, 150, 200, and 250 responses

5.1. Item Pool

The item pool consisted of a random sample of 250 retired items from a real-world testing program calibrated under the 3PL model. In addition, 50 items were sampled to serve as field-test items. Figure 1 shows the distribution of the operational item parameters in the pool. The plot for the b_i parameters reveals a pool that was slightly on the more difficult side with a mean equal to 0.260 and a standard deviation of 1.267. As the items were originally calibrated using the method of maximum marginal likelihood (MML) estimation, their posterior distributions were taken to be normals with the MML estimates and their standard errors as mean and standard deviation, respectively. From each of the distributions, 500 values were randomly drawn and stored in the system to represent the parameters of the operational items during testing.

5.2. Simulated Adaptive Testing

The adaptive test consisted of a total of 25 items, 22 of which were operational and three field-test items. As it was unnecessary to hide field-test items from simulated examinees in this study, for practical purposes their position was chosen to be the 21st, 23rd and 25th in the test for each examinee. The test had eight different content specifications modeled as categorical constraints. No quantitative constraints were required, but otherwise the shadow-test model was entirely similar to (23)–(33). The simulated examinees had ability parameter values randomly sampled from the population served by the testing program, which was estimated to be $N(0.26, 1.4^2)$.

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

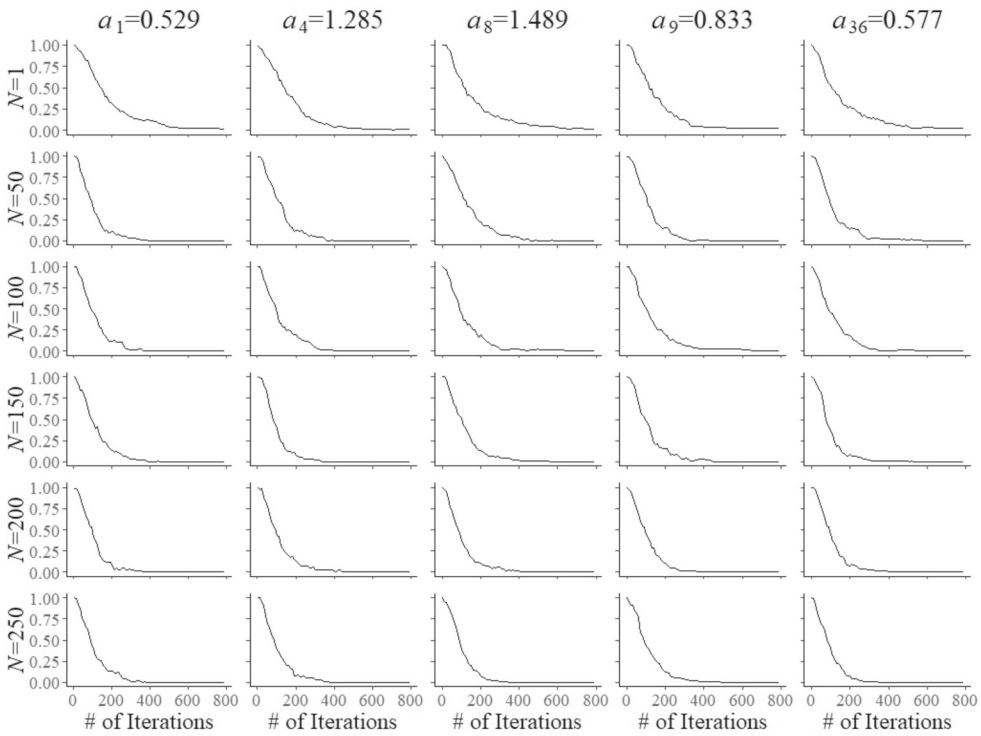


FIGURE 5.
Percentage of replications required to meet the convergence criterion of $\sqrt{\hat{R}} < 1.1$ as a function of the number of iterations for field test parameters a_f of five arbitrary items after $N = 1$, 50, 10, 50, 150, 200, and 250 responses

The Gibbs sampler for the ability parameter updates consisted of resampling of the 500 posterior draws for the operational item parameters in the system and the MH step in (7)–(11). Burn-in was for 250 iterations, while 500 independent post burn-in draws were saved for the next steps. The sample of 500 was obtained by thinning the post burn-in part of the chain by a factor of ten. These choices have been shown to be conservative for adaptive testing under the 3PL model (van der Linden and Ren 2020). The algorithm started with an initial prior distribution of $N(0.26, 1.68^2)$ for the ability parameter for each of the simulated examinees.

5.3. *Item-Calibration Algorithm*

The algorithm for item calibration was exactly the version of the Gibbs sampler above with the temporary reparameterization of the field-test parameters and subsequent use of the criterion of D_s-optimality in (22). The initial prior distributions of the field-test parameters were normals for b_i, $\ln(a_i)$, and $\text{logit}(c_i)$. As their parameters, the means and SDs for these quantities for the operational item pool were chosen, which were equal to 0.260 (1.520), −0.396 (0.582), and −1.414 (0.528), respectively. Because the D_s-criterion does not discriminate between items with a common initial prior distribution for their parameters, the simulation started with random assignment of each of the items to five examinees. The calibration process for an item was stopped as soon as it reached the criterion of $N = 250$, 500 or 1,000 examinees. For the 3PL model, these calibration sample sizes are low relative to those typically used for off-line item calibration, especially the two sizes of 250 and 500. Once an item was calibrated, it was treated as an operational item but with a large penalty as weight for its decision variable in the objective function in the shadow-test model. As a

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

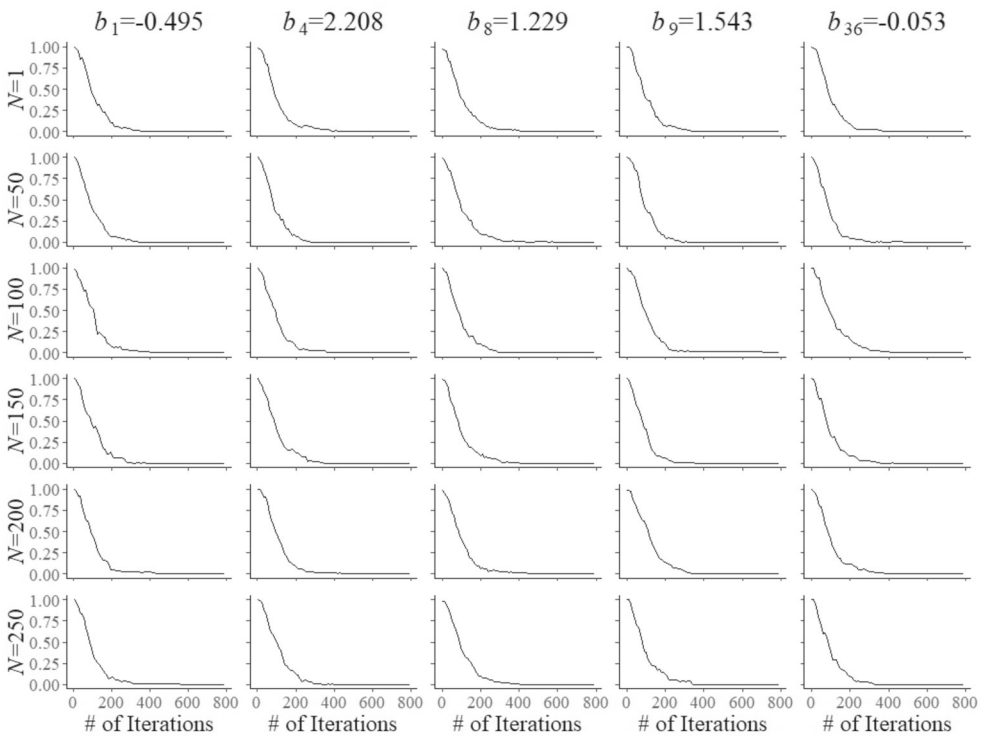


FIGURE 6.
Percentage of replications required to meet the convergence criterion of $\sqrt{\hat{R}} < 1.1$ as a function of the number of iterations for field test parameters b_f of five arbitrary items after $N = 1, 50, 10, 50, 150, 200$, and 250 responses

consequence, the model tried to avoid selecting the item but at the same time kept the impact of its presence on the relative severeness of the constraints in the model constant during the simulation.

5.4. Simulation Environment

All simulations were run on an i5-6300U 2.4GHz CPU with 16GB of RAM. The simulation framework was written in *R* with the Gibbs samplers coded in *Java*. The shadow tests were assembled making calls to a MIP solver installed on the same machine. Various open-source and commercial solvers are available. Open-source solvers as *lpSolve* and *CBC* (Forrest and Lougee-Heimer 2015) are free to use and can be easily integrated into most simulation frameworks. But commercial solvers as *FICO Xpress* and *IBM CPLEX* allow for significant reduction of solving time without any sacrifice of solution quality. For instance, the *FICO Xpress* solver used in the current simulations has built-in pre-solve algorithms and cut strategies producing solutions to typical problems 10 to 20 times faster than open-source solvers (Meindl and Templ 2012). The following settings were used for the main parameters of this solver: MIPRELSTOP = 1.0E−04, MIPABSSTOP = 0.0, MIPTOL = 5.0E−06, and FEASTOL: 1.0E−06. For a full list of parameters, including their definitions and default values, see the *FICO XPRESS Optimizer Reference Manual*.

5.5. Results

Convergence of the Markov chains for the updates of the field-test parameters is illustrated by the trace plots in Figs. 2, 3 and 4. The plots are for five arbitrary items with different true parameter

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

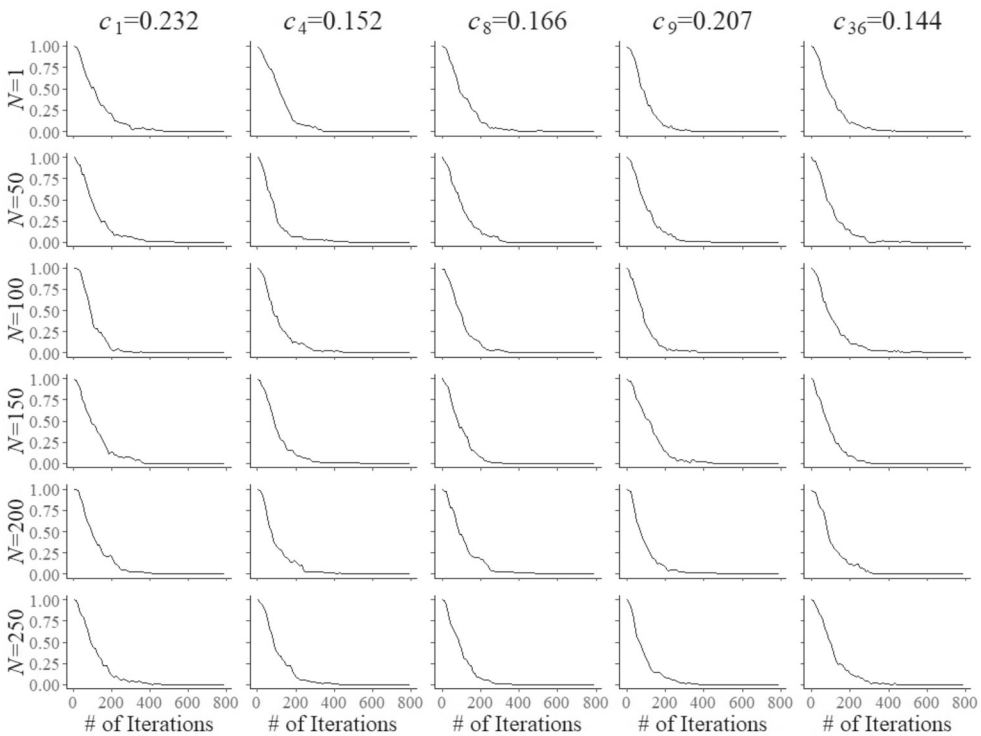


FIGURE 7.
Percentage of replications required to meet the convergence criterion of $\sqrt{\hat{R}} < 1.1$ as a function of the number of iterations for field test parameters c_f of five arbitrary items after $N = 1$, 50, 10, 50, 150, 200, and 250 responses

values after the responses by N = 1, 50, 100, 150, 200, and 250 examinees. Of course, after a single examinee the behavior of the chains was rather wild. But with increasing numbers, just as required, they quickly became regular with smaller and smaller posterior uncertainty located at the true values of the parameters.

Convergence of the chains was assessed more analytically using the well-known Gelman-Rubin (1992) diagnostic. In a separate set of simulations, four parallel chains were started from values randomly sampled from the tails of the initial prior distribution for each of the parameters, two from below its mean minus 2SD and the other two from above the mean plus 2SD. The chains were considered to be mixed once a potential scale reduction factor of $\sqrt{\hat{R}} < 1.1$ was reached. The percentages of replications required to meet the criterion as a function of burn-in length for the same five items and numbers of examinees are displayed in Figs. 5, 6 and 7. The results are quite robust. With the exception of the update after $N = 1$ examinee, all plots are remarkably similar, even across the three different types of parameters. Just to remain on the safe side, it was decided to run the main simulations with a burn-in length of 800 iterations for the updates of each of the field-test parameters.

A second set of simulations was run to estimate the autocorrelation of the Markov chains as a function of the lag size during the updates of the field-test parameters. Its results were necessary to obtain samples of independent post burn-in draws. Figures 8, 9 and 10 illustrate the functions estimated from 1,000 replications for the same items and numbers of examinees as in the previous figures. The choice of a lag size of $l = 20$ iterations was large enough to meet the criterion of autocorrelation of $\rho_l < 0.1$ for all simulated conditions. The lag size was then used to thin the

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

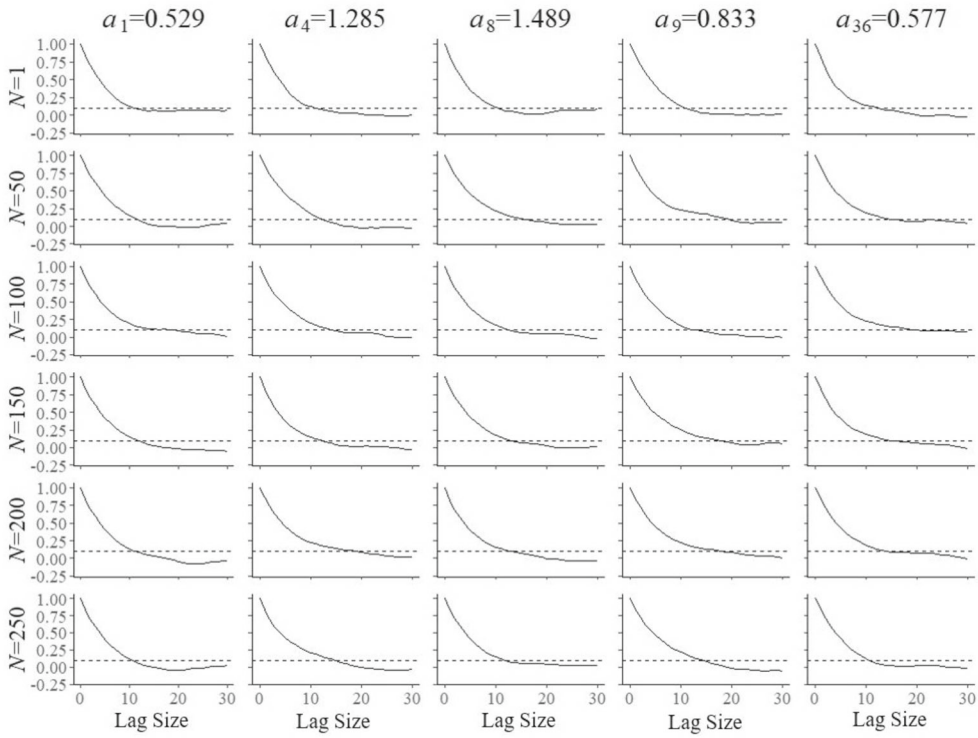


FIGURE 8.
Autocorrelation in the Markov chains as a function of the lag size for field test parameter a_f of five arbitrary items after $N = 1$, 50, 10, 50, 150, 200, and 250 responses

post burn-in chain to obtain samples of $S = 500$ independent draws for the selection of the next item and parameter update. The Monte Carlo error added to standard error of the posterior mean is known to be equal to a factor of $\sqrt{1 + 1/S} \times 100\%$ (Gelman et al. 2014, Sect. 11.5), which for our choice of $S = 500$ amounts to a relative error as small as 0.1%.

The main simulation conducted to evaluate the complete shadow-test approach was for 12, 000 simulated examinees with ability parameters sampled from $N(0.26, 1.4^2)$ and the choice of burn-in length and number of independent post burn-in draws equal to 250 and 500 for the update of the ability parameters and 800 and 500 for the update of the field-test parameters, respectively. Figures 11, 12 and 13 show the EAP estimates of the field-test parameters when the three different calibration sample sizes were reached plotted against their true values. The results for the b_f and c_f parameters for a sample size of $N = 250$ were already close to those typically seen in fixed-form item calibration for the 3PL model. In fact, the increase in sample size from $N = 250$ to $N = 1, 000$ did hardly have any impact on the accuracy of the estimates for these two parameters. The increase was welcome for the a_f parameters, however. The larger biases for the smallest sample size for these parameters disappeared with the increase of it, with the exception of a few items with higher true values for the parameter for which the decrease was only minor. The authors checked the simulated runs for these items carefully but have been unable to detect any irregularities or factors that could explain the exceptions. They may just have been the result of the randomness inherent in the procedure.

The criterion of D_s-optimality is known to be discriminating with respect to the assignment of items to examinees. Generally, it tends to prefer a few items in the pool, moving to the next

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

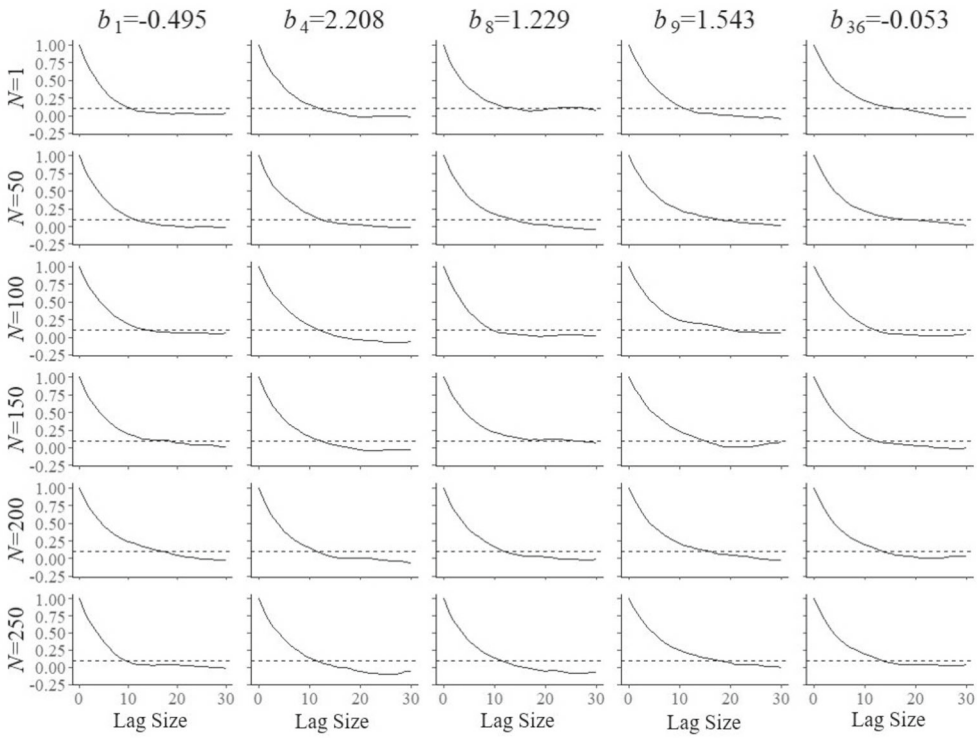


FIGURE 9.
Autocorrelation in the Markov chains as a function of the lag size for field test parameter b_f of five arbitrary items after $N = 1$, 50, 10, 50, 150, 200, and 250 responses

few only when their calibration is completed (van der Linden and Ren 2015). The behavior was observed in almost pure form in the current study. Figure 12 displays the number of items that completed the calibration as a function of number of simulated examinees. The functions were nearly perfectly linear for each of the three sample sizes until the maximum number of items was reached. These results point at an item-calibration approach that produces a constant stream of new items ready for operational use by the testing program right from its launch.

5.6. Runtimes

The distributions of the times for the Gibbs sampler to update the ability and field-test parameters and for the calls to the MIP solver with subsequent selection of the next item in the main study with 12,000 examinees are shown in Fig. 13. Their averages (standard deviations) were equal to 0.012 (0.003), 0.026 (0.003), and 0.053 (0.007) s, respectively. These results demonstrate appropriateness of the approach for use in real-world testing.

6. Discussion

Key features of the proposed approach to real-time item calibration are embedding of the field-test items in operational adaptive testing in random positions near the end of the test, representation of each of the response model parameters by short vectors of draws from their posterior distributions instead of fixed point estimates, sequential updating of these vector of posterior draws

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

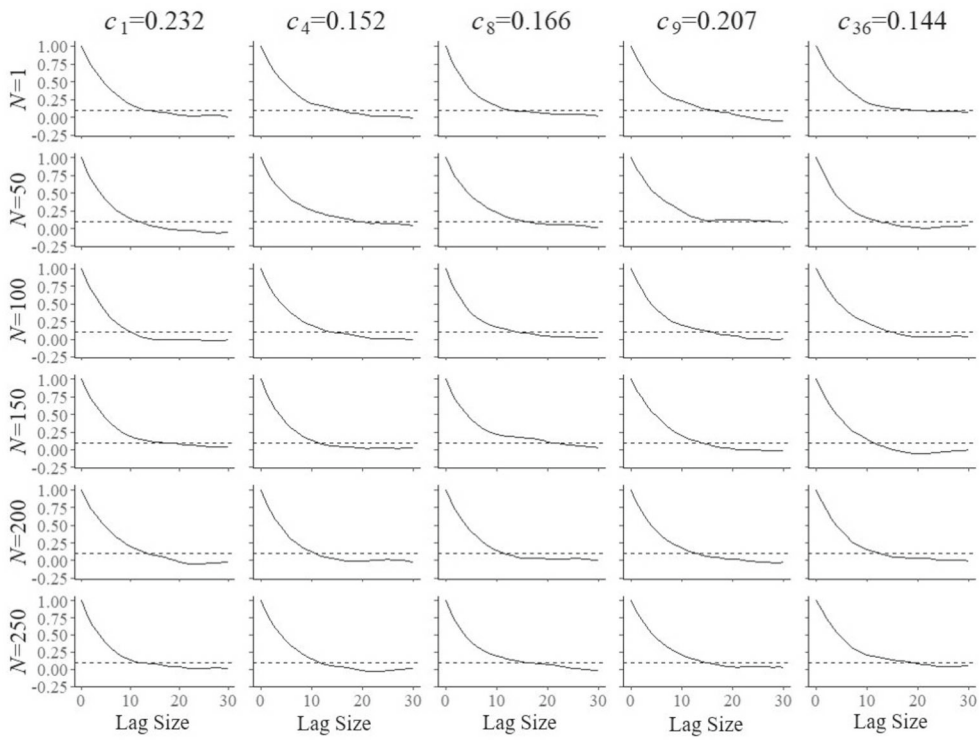


FIGURE 10.
Autocorrelation in the Markov chains as a function of the lag size for field test parameter c_f of five arbitrary items after $N = 1$, 50, 10, 50, 150, 200, and 250 responses

after each new response by two special versions of the Gibbs sampler, use of a well-known optimal design criterion both for the selection of the operational and field-test items, and maintenance of the integrity of the adaptive testing program through a shadow-test approach designed to keep all non-statistical parameters of the program within their bounds. Even for the rather conservative choice of settings for the two Gibbs samplers in our main study, the observed runtimes were fully comparable with those for conventional adaptive testing without any item calibration.

The approach delivers a constant stream of new items ready to replace obsolete items in the pool when their parameter updates have reached the desired degree of statistical precision. The fact that the parameters of these items are estimated directly on the scales in use for the item pool looks especially valuable. It prevents the necessity of separate parameter linking studies after item calibration and promises the best possible scale maintenance through the use of the entire item pool as anchor rather than a few selected items. The expected gains in item parameter stability relative to conventional linking studies have not been evaluated yet and deserve further study.

The current study carefully optimized the settings of the two Gibbs samplers for their application in the proposed shadow-test approach to adaptive testing. Though the settings for their main application in the study reported above were chosen to be conservative and the same choice has been found to work equally well in other applications, including adaptive testing from real-world pools of items calibrated under different polytomous IRT models (Ren et al. 2020), the authors do not claim universal validity. Critical factors not varied in the present study include the composition of the item pool, distribution of the item parameters, the set of test specifications, as well as the length of the test, each of which still has an unknown impact on the behavior of the proposed

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

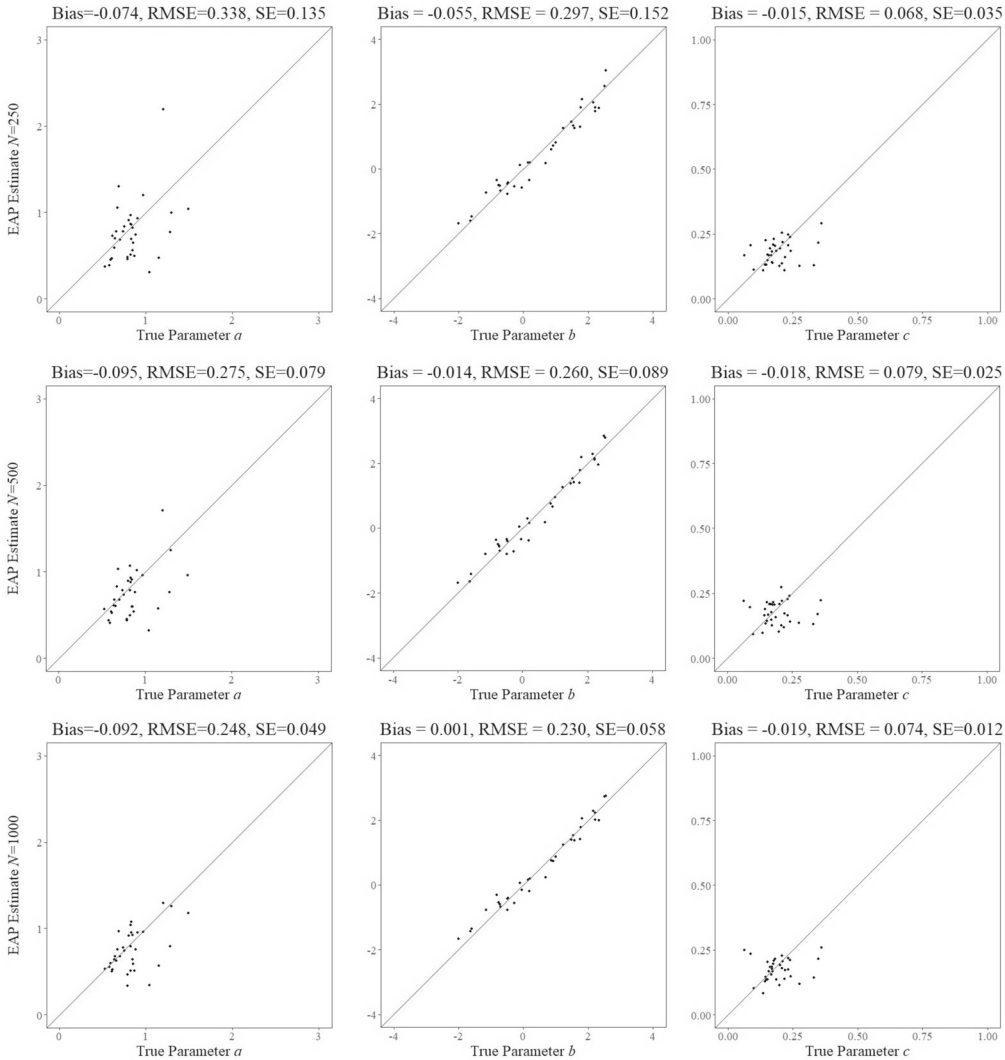


FIGURE 11.
Posterior means versus true values of all field test parameter after $N = 250$, 500, and 1,000 responses

algorithms. Testing programs interested in an application are recommended to run an additional simulation for their specific environment using our settings as point of departure.

The authors also are aware of fact that the linear results for the speed of item calibration demonstrated in Fig. 12 do not automatically hold for testing programs with continuous item writing and calibration rather than one fully calibrated batch of new items at a time as in the current study. As pointed out by Ren et al. (2017), it is not unlikely for a more continuous environment to find items with less than optimal statistical features being permanently dominated by some of the new items added to the pool. If so, they never complete their calibration, something especially undesirable when they belong to categories in the item pool that have become scarce. The issue brings us to the wider topic of item-pool management and its relation to field-testing of new items. The proposed shadow-test approach offers unique tools that support active management. One option to increase the speed of calibration of items in badly needed categories is using the

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

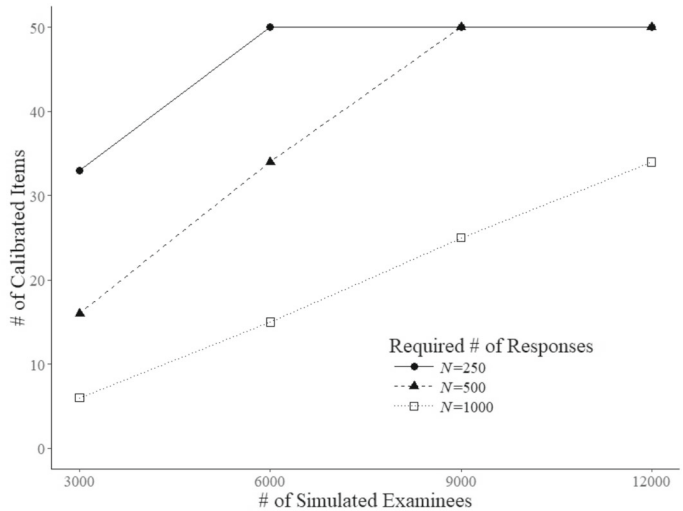


FIGURE 12.
Number of field-test items calibrated as a function of the number of simulated examinees after $N = 250$, 500, and 1,000 responses

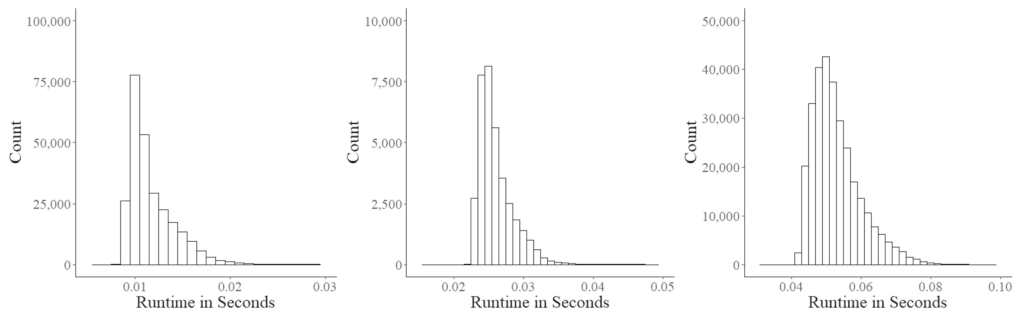


FIGURE 13.
Runtime distributions for the updates of ability parameter θ (left panel), updates of field-test parameters $\boldsymbol{\xi}_f$ (middle panel), and calls to the MIP solver with subsequent selection of the next item (right panel) in the main study (12,000 simulated examinees)

constraints in (30) to set lower bounds on their presence in the shadow tests. If the reason for lower than expected speed is less favorable statistical features of the field-test items, it is possible to put extra weight on their coefficients in the objective function of the model in (23). Both options can be exercised temporarily when monitoring the developments in the item pool in real time.

As stopping criterion, our study used fixed numbers of responses per item. An alternative criterion is a fixed level of posterior accuracy for the items. As both criteria are monotonically related for each item, we do not expect much different results for the latter. In fact, we expect sensible testing programs to monitor the posterior behavior of the field-test items until a prespecified level has been reached while capping their number of administrations to avoid occasional overexposure. When in this monitoring mode, they may even wish to decide stopping the calibration process for items beginning to show less than satisfactory fit to the response model for their program before either criterion is reached.

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Berger, M. P. F. (2018). Item calibration designs. In W. J. van der Linden (Ed.), *Handbook of item response theory. Volume 3: Applications* (pp. 3–20). Boca Raton, FL: Chapman & Hall/CRC.

CBC User Guide. Retrieved from https://www.coin-or.org/Cbc/.

Chang, H.-H., & Ying, Z. (2009). Nonlinear sequential design for logistic item response models with applications to computerized adaptive tests. *The Annals of Statistics*, *37*, 1466–1488.

Chen, D.-S., Batson, R. G., & Dang, Y. (2010). *Applied integer programming*. New York: Wiley.

FICO Xpress Optimizer Reference Manual. Retrieved from https://www.fico.com/fico-xpress-optimization/docs/latest/solver/optimizer/HTML/GUID-3BEAAE64-B07F-302C-B880-A11C2C4AF4F6.html.

FICO. *Xpress optimization*. Retrieved from https://www.fico.com/en/products/fico-xpress-optimization.

Forrest, J., & Lougee-Heimer, R. (2015). *CBC user guide*. Retrieved from https://www.coin-or.org/Cbc/.

Gelman, A., Carlin, J. B., Stern, H., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Holling, H., & Schwabe, R. (2018). Statistical optimal design theory. In W. J. van der Linden (Ed.), *Handbook of item response theory. Volume 2: Statistical tools* (pp. 313–339). Boca Raton, FL: Chapman & Hall/CRC.

IBM. *CPLEX optimizer*. Retrieved from https://www.ibm.com/analytics/cplex-optimizer.

lp_solve referenceguide. Retrieved from http://lpsolve.sourceforge.net/5.5/ [2] C.

Meindl, B., & Templ, M. (2012). Analysis of commercial and free and open source solvers for linear optimization problems. *Eurostat and Statistics Netherlands within the project ESSnet on common tools and harmonized methodology for SDC in the ESS*.

Ren, H., Choi, S. W., & van der Linden, W. J. (2020). Bayesian adaptive testing with polytomous items. *Behaviormetrika*. https://doi.org/10.1007/s11336-020-09703-8.

Ren, H., van der Linden, W. J., & Diao, Q. (2017). Continuous online item calibration: Parameter recovery and item calibration. *Psychometrika*, *82*, 498–522.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer.

van der Linden, W. J. (2018). Adaptive testing. In W. J. van der Linden (Ed.), *Handbook of item response theory. Volume 3: Applications* (pp. 197–227). Boca Raton, FL: Chapman & Hall/CRC.

van der Linden, W. J., & Choi, S. W. (2019). Improving item-exposure control in adaptive testing. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12254.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, *22*, 259–270.

van der Linden, W. J., & Ren, H. (2015). Optimal Bayesian adaptive design for item calibration. *Psychometrika*, *80*, 263–288.

van der Linden, W. J., & Ren, H. (2020). A fast and simple algorithm for Bayesian adaptive testing. *Journal of Educational and Behavioral Statistics*, *45*, 58–85.

Williams, H. P. (2013). *Model building in mathematical programming* (5th ed.). New York: Wiley.

Manuscript Received: 18 JUN 2019
Final Version Received: 19 DEC 2019
Published Online Date: 17 JUN 2020

https://doi.org/10.1007/s11336-020-09703-8 Published online by Cambridge University Press