




A Two-Level Adaptive Test Battery

Wim J. van der Linden

University of Twente

Luping Niu 

Seung W. Choi

The University of Texas at Austin

A test battery with two different levels of adaptation is presented: a within-subtest level for the selection of the items in the subtests and a between-subtest level to move from one subtest to the next. The battery runs on a two-level model consisting of a regular response model for each of the subtests extended with a second level for the joint distribution of their abilities. The presentation of the model is followed by an optimized MCMC algorithm to update the posterior distribution of each of its ability parameters, select the items to Bayesian optimality, and adaptively move from one subtest to the next. Thanks to extremely rapid convergence of the Markov chain and simple posterior calculations, the algorithm can be used in real-world applications without any noticeable latency. Finally, an empirical study with a battery of short diagnostic subtests is shown to yield score accuracies close to traditional one-level adaptive testing with subtests of double lengths.

Keywords: *ability estimation; adaptive testing; Bayesian optimality; Gibbs sampler; item response models; MCMC algorithm*

Introduction

Most educational and psychological testing programs are organized as a coherent battery of subtests. For example, for programs of diagnostic testing for instructional purposes, admission or vocational guidance decisions, and large-scale assessments of educational progress, reporting profiles of scores to stakeholders is much more informative than single scores summarizing an instructional need, a recommended decision, or progress made in a school district or country. A recent example from neurocognitive assessment is reported in Moore et al. (2023).

There also exist valuable psychometric benefits related to the use of test batteries. Obviously, a unidimensional response model is much more likely to fit each of a set of homogeneous subdomains of test items than the more

heterogeneous pool resulting from their aggregation. In addition, as scores inferred from subtests of a battery typically correlate highly, when dealing with one of the subtests, we have the opportunity to import data collected for the other subtests as powerful collateral information. The test batteries in this article are assumed to be such batteries of homogenous, unidimensional tests. As our focus is on adaptive testing, the assumption implies an item pool with the items for each subtest selected from a unique subset in the pool.

Test batteries are typically administered in fixed times slots, the presence of which imposes a difficult dilemma with respect to the number of subtests and their length. Generally, the greater the number, the richer the score profile. At the same time though, a greater number of subtests means fewer items per subtest and/or greater degree of speededness for some of them, the former implying lower accuracy of the score profile, the latter introducing bias with respect to its intended underlying abilities. No wonder the earliest applications of adaptive testing were attempts to use its efficiency to improve the design of test batteries. Since Lord's (1980, sect. 10.7) discovery of an adaptive version of the *PSAT* being at least twice as informative at almost all ability levels as a conventional version of the same length, his result has served as a rule of thumb for the efficiency of adaptive testing in the testing industry. By the same rule, adaptive testing could be used to increase the number of subtests in a battery by a factor of two without sacrificing any accuracy of the test scores, reduce the administration time by the same factor, or for a combination of both. Early pioneers exploring these new opportunities include Brown and Weiss (1977) and Gialluca and Weiss (1979).

The previous rule refers to the typical adaptive test starting with the selection of an item somewhere in the middle of the ability scale. However, the dilemma between the number of subtests and their score accuracy can be further relaxed if, instead, we initialize each next subtest using collateral information from other subtests in the battery. As already noted, as test batteries tend to have high intercorrelation between their subtests, we could do so with high precision. In fact, we could go even one step further and select the subtests for each test taker *adaptively*: That is, rather than administering all subtests in a fixed, arbitrary order, each with a fixed choice of the initial ability estimate, choose the next subtest with the currently best initialization of the test taker's ability among all remaining subtests. The process then implies two distinct levels of adaptation: a within-subtest level for the selection of the items from each subpool and a between-subtest level when moving from one subpool to the next.

The additional level of adaptation can be used to develop large batteries of short diagnostic subtests (each of 5–7 items, say), for instance, to monitor learning progress in education. As their improved initialization immediately brings the selected items close to the true abilities of the test takers, only a few more items are required to finish each subtest. The current research was motivated by the desire to replace current diagnostic testing with its typical subscoreing of long heterogeneous fixed tests with such batteries of a large number of short adaptive tests.

The idea of a two-level adaptive test battery was already proposed in van der Linden (2010). The proposal consisted of a battery run on a regular response model for each of the subtests extended with a second level for the joint distribution of their abilities. However, its statistical treatment was rather ad hoc. All item parameters and second-level ability parameters were assumed to be equal to point estimates obtained during earlier item pool calibration. But as known for traditional one-level adaptive testing, the treatment of point estimates as known parameters results in overoptimistic estimates of the test taker's ability parameter together with less than optimal item selection due to capitalization on item parameter error (e.g., Cheng et al, 2015; Patton et al, 2013; van der Linden & Glas, 2000). In the context of two-level adaptive testing, not only the within-subtest level of item selection is impacted by item parameter error, but the same can be expected to happen during the transition from one subtest to the next. The overestimation of the information about the first-level ability parameters from the preceding subtests and the second-level parameters for the ability structure is then likely to result in less than optimal selection of the next subtest as well.

Another necessary improvement is with respect to the final subtest scores reported in van der Linden (2000). As each next subtest profited from a greater number of responses already collected from the test taker and thus began with more information about the next ability parameter, the gains of score accuracy for a subtest were higher, the later it was administered to the test taker.

Finally, the earlier proposal was based on an ad hoc procedure to compute all necessary integrals using a single random sample of ability values drawn from the multivariate ability distribution with its means and (co)variances set equal to points estimates collected prior to the test. As the integrals were with respect to the conditional distributions of the current ability parameter given all possible combinations of the ability parameters for the preceding subtests, the necessary sample size to produce the accurate estimates of these continuous distributions quickly becomes prohibitive for test batteries with larger numbers of subtests.

The current proposal is based on a fully Bayesian approach. All first- and second-level parameters are assumed to be known only through their posterior distributions. The subpools and items are selected based on posterior distributions of the ability parameters permanently updated during testing, avoiding the danger of capitalization on parameter error inherent in adaptive testing based on point estimates. Also, computationally, rather than a large single sample from the second-level ability distribution, the updates are obtained locally from a rapidly converging Gibbs sampler. In addition, the new approach is extended with an adjustment that removes the unbalance between the scores on the earlier and later subtests in the battery scores, making each final score for an earlier subtest as informative as the score for the final subtest. Finally, as discussed at the end of this article, the combination of a two-level adaptive testing model with the fully Bayesian approach allows for several extensions and generalizations of adaptive testing, including such practical options as continued updating of the model

parameters during operational testing or even continuous field testing and calibration of new items.

In the next sections, we first review the two-level response model for the adaptive test battery and then introduce our Bayesian approach to ability parameter updating, item selection, and the transition from one subpool to the next. Using the output from the proposed Gibbs sampler, the computational expressions for the optimization of each of these steps are presented. The practical feasibility of the approach is demonstrated for an extensive study with simulated test takers for a pool of items for a real-world adaptive test battery.

Two-Level Model

Each of the subpools of items $h = 1, \dots, H$ is assumed to measure a distinct ability. In addition, we use $i_h = 1_h, \dots, I_h$ to denote the i th item in subpool h . The size of subpool h is thus equal to I_h . The necessary model equations are at two levels: a lower level for the adaptive selection of the items within the subtests and a higher level for the adaptive sequencing of the subtests.

The lower level equations define the well-known three-parameter logistic (3PL) response functions:

$$\Pr\{U_{i_h} = 1\} \equiv c_{i_h} + (1 - c_{i_h}) \frac{\exp[a_{i_h}(\theta_h - b_{i_h})]}{1 + \exp[a_{i_h}(\theta_h - b_{i_h})]}, \quad h = 1, \dots, H, \quad (1)$$

where $b_{i_h} \in \mathbb{R}$ and $a_{i_h} \in \mathbb{R}^+$ can be interpreted as parameters for the difficulty and discriminating power of item i from subpool h , respectively, and $c_{i_h} \in (0, 1]$ as the probability of a correct response to the item resulting from random guessing. For convenience, we will use vector notation $\boldsymbol{\xi}_{i_h} \equiv (a_{i_h}, b_{i_h}, c_{i_h})$ to denote the parameters of item i_h . The choice of the 3PL model is because of its popularity only. Any other response model with separate item and test taker parameter suitable for adaptive testing could have been chosen instead.

At the higher level, the ability structure in the population of the test takers is supposed to follow a multivariate normal density

$$f(\theta_1, \dots, \theta_H) \equiv MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (2)$$

with mean vector $\boldsymbol{\mu}$ and $H \times H$ covariance matrix $\boldsymbol{\Sigma}$. The multivariate structure will be needed for its marginal distributions to select the first subtest as well as each of its possible conditional distributions to continue with the selection of the subsequent subtests. As for these conditional distributions, let \mathbf{c} denote the vector of indices of the abilities already tested and θ_h the ability of current interest. From the general theory of multivariate normal distributions (e.g., Eaton, 1983, pp. 116–117), we know that $f(\theta_h | \boldsymbol{\theta}_c)$ is normal with conditional mean and variance

$$\boldsymbol{\mu}_{h|c} = \boldsymbol{\mu}_h + \boldsymbol{\Sigma}_{hc} \boldsymbol{\Sigma}_{cc}^{-1} (\boldsymbol{\theta}_c - \boldsymbol{\mu}_c), \quad (3)$$

and

$$\sigma_{h|c}^2 = \sigma_h^2 - \Sigma_{hc} \Sigma_{cc}^{-1} \Sigma'_{hc}, \quad (4)$$

where μ_c and Σ_{cc} are the mean vector and covariance matrix of conditioning abilities θ_c , Σ_{hc} is the vector with the covariances between θ_h and θ_c , and μ_h and σ_h^2 are the marginal mean and variance of θ_h .

The items in each subpool are assumed to have been calibrated prior to operational adaptive testing. The details of the Bayesian procedure used by the authors in their empirical example are provided in the following section. For applications unlikely to have the multivariate normal structure in Equation 2, it is possible to introduce transformations of the abilities deviating from normality with back transformation of the final scores. Another straightforward option, seamlessly fitting the approach below but less attractive from a statistical point of view, is discussed at the end of this article.

Our goal for the test battery is optimal estimation of the score profiles $(\theta_1, \dots, \theta_H)$ for the each of the test takers. If the interest would be in a summary of these profiles only, a linear combination of the scores with policy weights specified by the owner/stakeholders of the program could be used. If so, alternatively, the third-level model by de la Torre and Song (2009) with a latent overall ability parameter for each test taker and each of the abilities θ_h regressed on the overall parameter is available. However, our current goal requires the assumption of the two levels in Equations 1 and 2 only.

Bayesian Approach

The approach is fully Bayesian accounting for the uncertainty about all parameters in Equations 1 and 2. At each step in the approach, all prior distributions are empirical distributions.

The initial prior distributions for the ability parameters used for the selection of the first subpool are the marginal distributions of Equation 2 saved from item calibration. The distributions are updated in a sequential fashion during testing using a generalization of the Gibbs sampler introduced in van der Linden (2018) and further optimized by van der Linden and Ren (2020), Ren et al. (2020), and Niu and Choi (2022). The prior distributions for the ability parameters used for the selection of the later subpools are the posterior predictive distributions immediately available upon termination of the preceding subtest, whereas the prior distributions for the item and second-level model parameters are the posterior distributions obtained from their calibration prior to operational testing. As demonstrated by our examples in the following, though quite efficient, this Bayesian approach is computationally not more intensive than traditional one-level adaptive testing with point estimation of all parameters.

The proposed Gibbs sampler is introduced first allowing its output to be used in the presentation of the necessary computational expressions for subpool and item selection.

Gibbs Sampler

The algorithm for the updates of the ability parameters is based on the following ideas:

1. Rather than the usual point estimates of the item parameters, ability parameters, and means and covariances in Equations 1 and 2, short vectors of random draws from their last posterior distributions are stored in the testing system.
2. At each of the posterior updates of an ability parameter, a Gibbs sampler is used which cycles between
 - a. resampling the vector of draws saved for the item, mean, and covariance parameters; and
 - b. a Metropolis–Hastings (MH) step to sample the test taker’s ability parameter.
3. The MH steps for the ability parameter capitalize on the sequential nature of adaptive testing in the following way:
 - a. the proposal distribution is a normal centered at the value drawn at the preceding iteration step, and as variance, a rescaled version of the posterior variance saved from the previous update of the ability parameter and
 - b. the prior distribution is a normal with both the mean and variance saved from the previous update.
4. Upon termination of the MH steps, the existing vector of draws for the ability parameter in the system is overwritten by an appropriate selection from the current draws.

For the rescaling in Step 3a, Niu and Choi (2022) recommended a factor of 2.4 times the posterior standard deviation found to be most effective by them. As the posterior distributions of the item and second-level ability parameters do not depend on the data used to update the test taker’s ability parameter, their resampling replaces the sampling of the complete conditional distributions generally required for a Gibbs sampler. In fact, because of posterior independence, the sampler reduces to something known as an independence sampler (Gilks, Richardson & Spiegelhalter, 1996). Also, as the posterior distributions for these parameters are narrow and already converged, the Markov chain needs to converge for one ability parameter only, which occurs almost immediately.

The last posterior mean and variance of the ability parameter are always our best summary of the information about the parameter collected so far. Their use as prior distribution at each next update thus remains empirical. Besides, the proposal distribution used in the MH steps adapts itself in the sense of having a mean automatically converging to the true parameter value during testing and variance converging to zero. As the proposal distribution is symmetric, the acceptance probability for the candidate value drawn from it reduces to the

simple calculation of the product of the prior and model probability of the test taker's response given the current proposal for the value of the ability parameter. No other computational steps are required.

More formally, the MH step for the update of the posterior distribution of θ_h after the test taker's response to the $(k - 1)$ th item in subtest h is as follows: Let $f(\theta_h | \mathbf{u}_{k-1})$ denotes the posterior density of θ_h given response vector $\mathbf{u}_{k-1} \equiv (u_1, \dots, u_{k-1})$ and $f(\boldsymbol{\xi}_{i_{k-1}})$ the posterior density of the parameters of item i_{k-1} , where the dependency of $f(\boldsymbol{\xi}_{i_{k-1}})$ on the data used for item calibration is omitted for convenience. The prior distribution at this step is $N(\theta_h; \mu_{k-2}, \sigma_{k-2}^2)$, with μ_{k-2} and σ_{k-2}^2 the posterior mean and variance of θ_h from the previous update directly calculated from its Markov chain.

Using $t = 1, 2, \dots$ to denote the iterations of the sampler and $\theta_h^{(t-1)}$ for the draw at iteration $t - 1$, the necessary steps at iteration t are

1. Drawing a value from the posterior samples of item parameters $a_{i_{k-1}}$, $b_{i_{k-1}}$, and $c_{i_{k-1}}$ stored in the system.
2. Drawing a candidate value θ_h^* from the proposal density $N(\theta_h^{(t-1)}, \sigma_{k-2}^2)$.
3. Calculating the probability of acceptance

$$r = \frac{N(\theta_h^*; \mu_{k-2}, \sigma_{k-2}^2) \Pr\{U_{i_{k-1}} = u_{i_{k-1}} | \theta_h^*, \boldsymbol{\xi}_{i_{k-1}}^{(t)}\}}{N(\theta_h^{(t-1)}; \mu_{k-2}, \sigma_{k-2}^2) \Pr\{U_{i_{k-1}} = u_{i_{k-1}} | \theta_h^{(t-1)}, \boldsymbol{\xi}_{i_{k-1}}^{(t-1)}\}}. \quad (5)$$

4. Accepting $\theta_h^{(t)} = \theta_h^*$ with probability $\min\{r, 1\}$ but keeping $\theta_h^{(t)} = \theta_h^{(t-1)}$ otherwise.
5. Returning to Step 1.

As already hinted, the calculation of Equation 5 requires only the calculation of the product of the current prior density and the probability of the last observed response at candidate value θ_h^* in the numerator; the denominator was already calculated during the previous iteration. For applications in single-level adaptive testing, the Markov Monte Carlo chain (MCMC) chain produced by the algorithm has been shown to converge in less than a few hundred iterations. In addition, for autocorrelation negligible at lags larger than a small critical value l^* , the chain needs to be continued for no more than $100l^*$ iterations to save $S = 100$ independent draws for use in operational testing. At this point, the Monte Carlo error in the estimate of the posterior standard deviation of the ability parameter is not greater than 0.5% (Gelman et al., 2014, sect. 11.4-5). Continuing the chain to meet this criterion requires only a few extra milliseconds on a standard PC. For these and other details, the reader is referred to van der Linden and Ren (2020).

Fisher's Information

A crucial quantity in adaptive testing is Fisher's information in response U_i to item i about ability parameter θ_h , which is defined as

$$I_i(\theta_h; \xi_i) \equiv -\mathcal{E} \left[\frac{\partial^2}{\partial \theta_h^2} \ln l(\theta_h; U_i, \xi_i) \right], \quad (6)$$

$$= \frac{[p_i'(\theta_h; \xi_i)]^2}{p_i(\theta_h; \xi_i)[1 - p_i(\theta_h; \xi_i)]}, \quad (7)$$

where $l(\theta_h; U_i, \xi_i)$ is the likelihood statistic associated with response U_i and $p_i'(\theta_h; \xi_i)$ the first-order derivative of the probability function in Equation 1 with respect to θ_h while the expectation is taken across the distribution of the response. For the 3PL model, the expression is known to simplify to

$$I_i(\theta_h; \xi_i) = a_i^2 \frac{1 - p_i(\theta_h; \xi_i)}{p_i(\theta_h; \xi_i)} \left(\frac{p_i(\theta_h; \xi_i) - c_i}{1 - c_i} \right)^2. \quad (8)$$

The criterion for item selection is maximum posterior expected information, which, for a given response vector \mathbf{u} , is defined as

$$\int \int I_i(\theta_h; \xi_i) f(\theta_h | \mathbf{u}) f(\xi_i) d\theta_h d\xi_i, \quad (9)$$

where $f(\theta_h | \mathbf{u})$ and $f(\xi_i)$ are the posterior densities of θ_h and the parameters of item i , respectively.

We use superscripts $s = 1, \dots, S$ to denote draws from the posterior samples currently stored in the system. The criterion for the selection of item i in subtest h is simply calculated as

$$S^{-1} \sum_{s=1}^S I_i(\theta_h^{(s)}; \xi_i^{(s)}). \quad (10)$$

The use of equal numbers of draws is for notational convenience only. In case of unequal sample sizes, the average is efficiently calculated recycling smaller samples against the larger.

Selection of First Subpool

It may seem advantageous to administer the first test from the subpool with the item that has the greatest value of Equation 10 across all subpools, where the draws for θ_h are from the marginal distribution of Equation 2 stored in the system. However, this criterion has a serious disadvantage: A subpool may have only a limited number of excellent items, whereas the rest of them is inferior to larger numbers of items in some of the other subpools. The use of the criterion is then bound to lead to suboptimal selection of later items.

A solution that avoids this pitfall of capitalization on a few good items is to select the best full-size subtest from each subpool and pick the subpool with the best result among all of them. Let n_h denotes the size of subtest h and \mathcal{V}_{n_h} a set of

items of this size from subpool h . The appropriate criterion is the subpool with the greatest value of

$$\left\{ (n_h S)^{-1} \sum_{i_h \in V_{n_h}} \sum_{s=1}^S I_i(\theta_h^{(s)}; \boldsymbol{\xi}_i^{(s)}) \right\}, \quad (11)$$

among all possible sets V_{n_h} , where $\theta_h^{(s)}$ are random draws from the marginal distribution of θ_h in Equation 2. The draws are obtained in two steps by first drawing $\mu_h^{(s)}$ and $\sigma_h^{(s)}$ from the samples from their posterior distribution saved during item calibration followed by a draw from the normal distribution given the sampled values; that is, as

$$\theta_h^{(s)} \sim N(\mu_h^{(s)}, \sigma_h^{(s)}). \quad (12)$$

As indicated earlier, this article is based on research operating on the assumption of a larger battery of short subtests each from a homogenous subpool of items. But if the same methodology is applied to a battery with more heterogenous subpools, a second pitfall is possible. It may then be necessary to impose constraints on the selection of the items to balance the content of the subtest across all test takers. If so, just beginning a subtest with items that are statistically best for the test taker is likely to lead to later suboptimal selection because of the necessity to satisfy each of the constraints at the end of the test.

A solution that efficiently avoids both pitfalls is to use a shadow-test approach (STA) both for the selection of subpools and items. The first subtest is then the one with the best solution for the shadow-test model with Equation 11 as objective and a constraint set that controls both the length of the subtest and the content distribution of the items. For a brief review of the approach, see the Appendix.

Selection of Items From First Subpool

We relabel the subpools assigning $h = 1$ to the subpool from which the first test in the battery is administered. The first item in Subtest 1 is the one with the greatest value for the criterion in Equation 10, still with the draws from the marginal distribution of θ_1 substituted into it.

After each new response $u_{i_{k-1}}$, the Gibbs sampler with the acceptance probability in Equation 5 is run to update the posterior distribution of θ_1 . The next item is selected using Equation 10 again, but this time with the draws $\theta_1^{(s)}$ from the $(k - 1)$ th posterior update of θ_1 substituted into it. The draws from the update after the last item are saved for the selection of the second subpool.

Selection of Second Subpool

The second subtest is administered from one of the subpools $h = 2, \dots, H$. The posterior expected information in a response to item i from subpool h is

$$\int \int I_i(\theta_h; \boldsymbol{\xi}_i) f(\theta_h | \mathbf{u}_1) f(\boldsymbol{\xi}_i) d\theta_h d\boldsymbol{\xi}_i, \quad (13)$$

that is, the version of Equation 9 with $f(\theta_h | \mathbf{u}_1)$ replacing the marginal distribution of θ_h used as initial prior distribution when the first subpool was selected. Observe that

$$f(\theta_h | \mathbf{u}_1) = \int f(\theta_h | \theta_1) f(\theta_1 | \mathbf{u}_1) d\theta_1, \quad (14)$$

which defines $f(\theta_h | \mathbf{u}_1)$ as the predictive posterior density of θ_h given the responses from the first subtest.

Analogous to Equation 10, Equation 13 is calculated as

$$S^{-1} \sum_{s=1}^S I_i(\theta_h^{(s)} | \mathbf{u}_1^{(s)}, \boldsymbol{\xi}_i^{(s)}), \quad (15)$$

where $\theta_h^{(s)} | \mathbf{u}_1$ are random draws from the distribution in Equation 14. The draws are also obtained in two steps, by first drawing a value $\theta_1^{(s)}$ from $f(\theta_1 | \mathbf{u}_1)$ and then following with a draw from $f(\theta_h | \theta_1)$ given $\theta_1 = \theta_1^{(s)}$. The former are present in the system as draws saved from the last update of θ_1 . As for the latter, from Equations 3 and 4, we know that $f(\theta_h | \theta_1)$ is

$$N(\mu_h + \frac{\sigma_{h1}}{\sigma_h \sigma_1} (\theta_1 - \mu_1), 1 - (\frac{\sigma_{h1}}{\sigma_h \sigma_1})^2), h = 2, \dots, H. \quad (16)$$

Thus, the draws required for Equation 15 are obtained from

$$\theta_h^{(s)} | \mathbf{u}_1 \sim N(\mu_h^{(s)} + \frac{\sigma_{h1}^{(s)}}{\sigma_h^{(s)} \sigma_1^{(s)}} (\theta_1^{(s)} - \mu_1^{(s)}), 1 - (\frac{\sigma_{h1}^{(s)}}{\sigma_h^{(s)} \sigma_1^{(s)}})^2), s = 1, \dots, S, \quad (17)$$

with the second-level parameters resampled from their posterior distributions saved from the calibration of the item pool.

The second item pool is selected according to the criterion in Equation 11 with the conditional draws $\theta_h^{(s)} | \mathbf{u}_1$ replacing the draws from the marginal distribution used to select the first subpool. Both the improved location and smaller variance of the conditional relative to the marginal distribution are indicative of the increase in efficiency of the test battery due to the responses collected from the first subtest when selecting the second.

Selection of Items From Second Subpool

We now use $h = 2$ to denote the subpool used for the second subtest. The first item from this pool is the one with the greatest value for the criterion in Equation 10, still with $\theta_h^{(s)}$ replaced by $\theta_2^{(s)} | \mathbf{u}_1$. The next items from the subpool are selected using the Gibbs sampler to update the test taker's posterior

distribution of θ_2 , given the full response vector \mathbf{u}_1 and partial vector \mathbf{u}_2 . However, it is no longer necessary to condition explicitly on u_1 , as was required when selecting the second subpool. Each posterior distribution used as prior when selecting the next item already contains the accumulated information from the responses to all previous items administered to the test taker.

Selection of Subsequent Subpools and Items

The same procedure is continued to select subsequent subpools and items. The only necessary change for the selection of the next subpool is the extension of Equations 13 through 17 with an additional conditioning ability parameter representing the last subtest administered. To illustrate one more step, it is easy to verify from the general result in Equations 3 and 4 that, using correlations rather than covariances for notational convenience, for the selection of the third subpool, $f(\theta_h|\theta_1, \theta_2)$ has conditional mean and variance

$$\mu_{h|1,2} = \mu_h + \frac{1}{1 - \rho_{12}^2} (\rho_{h1} - \rho_{h2}\rho_{12})\theta_1 + (\rho_{h2} - \rho_{h1}\rho_{12})\theta_2, \quad (18)$$

and

$$\sigma_{h|1,2}^2 = \sigma_h^2 - \frac{1}{1 - \rho_{12}^2} (\rho_{h1}^2 + \rho_{h2}^2 - 2\rho_{h1}\rho_{h2}\rho_{12}), \quad (19)$$

respectively. Thus, analogous to Equation 17, we combine the draws $\theta_1^{(s)}$ and $\theta_2^{(s)}$ saved upon the completion of the first and second subtests along with $\rho_{h1}^{(s)}$, $\rho_{h2}^{(s)}$, and $\rho_{12}^{(s)}$, $s = 1, \dots, S$, to obtain the draws from the normal distribution of $f(\theta_h|\theta_1, \theta_2)$ necessary for the application of the next version of the criterion in Equation 11.

For larger numbers of subpools, the use of analytic expressions for the conditional means and variances derived directly from Equations 3 and 4 becomes less convenient. A more practical approach is then to use the fact that conditional variance $\sigma_{h|c}^2$ is the Schur complement of Σ_{cc} in the submatrix of Σ with the covariances between θ_h and the conditioning abilities $\theta_1, \dots, \theta_{h-1}$. The required $\sigma_{h|c}^2$ is the reciprocal of the h th diagonal element of the inverse of the submatrix, which is easily obtained using one of the standard routines for matrix inversion. The conditional mean should be calculated directly from Equation 3. The selection of the items from each of the subsequent subpools still amounts just to another application of the Gibbs sampler with the common mean and variance in the denominator of Equation 5 saved from the immediately preceding update of the ability parameter.

Final Subtest Scores

The procedure presented so far suggests a serious unbalance in the sense of earlier subtests necessarily profiting from a smaller number of preceding subtests and hence producing less accurate final scores than later subtests. The proposed correction for the unbalance is to recalculate the final scores for the subtests from the posterior distribution of each ability parameter given the test taker's complete set of responses to all subtests; that is,

$$f(\theta_h | \mathbf{u}_1, \dots, \mathbf{u}_H), \quad h = 1, \dots, H. \quad (20)$$

A straightforward way to sample Equation 20 for each θ_h is to rerun the complete battery reseeding the responses collected for each of the items administered to the test taker into the system. The only requirement is that subtest h should be in the last position; otherwise, the order of the other subtests is arbitrary. The approach is pragmatic in that it does not require any new computer code, only the code for the current Gibbs sampler to draw from the update of the posterior ability distribution after each of the reseeded responses.

An alternative approach is to redesign the Gibbs sampler to update the posterior distribution of each θ_h directly from the complete collection of responses by the test taker. This approach, recommended for application in large-scale operational testing, does require new code though along with a separate study to find the required burn-in, estimate the autocorrelation, and so on for the extended sampler.

The means and *SDs* of the draws from this final update of θ_h can be used to report the profile with all subtest scores along with their accuracies to the test taker.

Empirical Example

The goal of the empirical example was to demonstrate the practical feasibility of the current approach to two-level adaptive testing and give an impression of the gain in relative efficiency created by the introduction of the second level of adaptation under operational conditions.

Item Pool Calibration

The real-world test battery used in the example had four different subtests labeled here as Subtests 1 through 4. The item pool consisted of 150 items randomly sampled from an inventory of retired operational items for each of the subtests. The items had been extensively pretested and shown to have satisfactory fit to the 3PL model in Equation 1. Also, the ability parameters for the four subpools had been estimated to have empirical mean vector

$$\boldsymbol{\mu} = (-0.92, -1.04, -0.62, -0.96), \quad (21)$$

and covariance matrix

$$\Sigma = \begin{bmatrix} 1.02 & 0.63 & 0.85 & 0.78 \\ 0.63 & 0.80 & 0.64 & 0.71 \\ 0.85 & 0.64 & 1.21 & 0.89 \\ 0.78 & 0.71 & 0.89 & 1.18 \end{bmatrix}, \quad (22)$$

for a typical population of test takers. The estimates of all these parameters were used as their true values for the generation of response data, both for item pool calibration and during simulated adaptive testing.

The items had been calibrated previously using one of the standard computer programs from the maximum-likelihood tradition. But as samples from the posterior distributions for all parameters in the two-level model in Equations 1 and 2 were needed to simulate the adaptive testing administrations, it was decided to re-estimate all parameters in a fully Bayesian version. (An alternative would have been to take the maximum-likelihood estimates (MLEs) together with their estimated standard error and sample the parameters assuming asymptotic normality. But Bayesian estimation was preferred because of the small-sample validity of its posterior distributions.) All parameters were estimated jointly in a Bayesian fashion from the response data of $N = 1,000$ test takers generated for each item, using a Gibbs sampler implemented for the two level-model in *JAGS* (Plummer, 2017). The prior distributions for the items parameters were chosen to be $a_i \sim N(1, .5^2)I(a_i > 0)$, $b_i \sim N(0, 2^2)$, and $c_i \sim Beta(2, 5)$. The prior distribution for the second-level parameters was the conjugate normal-inverse-Wishart, with as hyperparameters the empirical mean vector and covariance matrix in Equations 21 and 22 and $k = 20$ degrees of freedom. The other settings for the sampler were a burn-in of 5,000 iterations and thinning of the remaining portion of the chain by a factor of 500. The posterior samples saved for each of the parameters for use in adaptive testing consisted of 500 independent draws. Each of these choices was made based on the results from an extensive study of the sampler to optimize its use for more general two-level IRT applications.

Figure 1 shows the scatterplots of the posterior means (expected a posteriori or EAP estimates) against the true values of the item parameters for the four subpools. The average root mean squared errors (RMSEs) for the item parameters are shown in Table 1. For second-level parameters μ and Σ , the average RMSE was equal to 0.011 and 0.129, respectively. These results reveal enough remaining parameter uncertainty to motivate the current Bayesian approach to adaptive testing.

Simulation Conditions

The main conditions in the simulation study were:

1. subtest length of 5 versus 10 items;
2. two-level versus one-level adaptive testing.

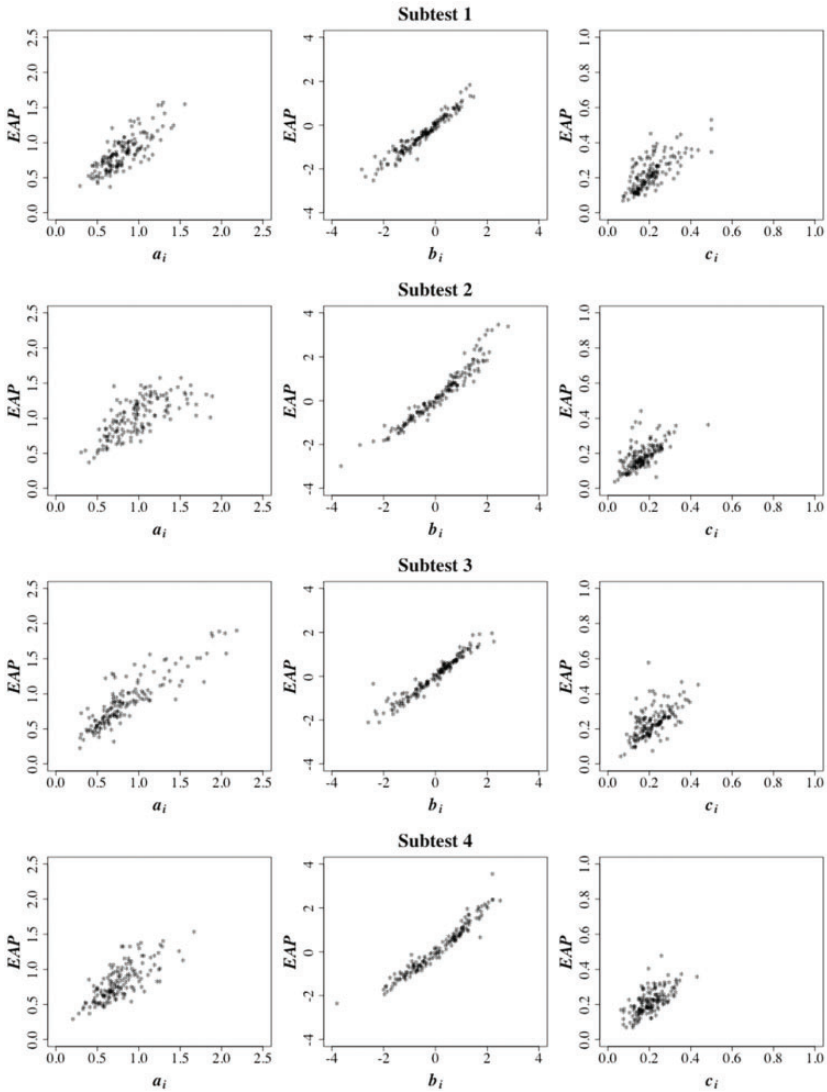


FIGURE 1. Scatterplots of the posterior means of the a_i , b_i , and c_i parameters against their simulated true values for each of the four subpools.

Each of the four combinations of conditions was simulated for a total of 5,000 test takers with ability parameters θ_h sampled from the multivariate normal distribution with the mean vector and correlation matrix in Equations 21 and 22. To be able to report accurate results for the tails of the ability distribution in

TABLE 1.

Average Root Mean Squared Errors for the a_i , b_i , and c_i Parameters for Each of the Four Subpools

Subpool	a_i	b_i	c_i
1	.250	.275	.056
2	.350	.291	.049
3	.341	.291	.061
4	.358	.265	.450

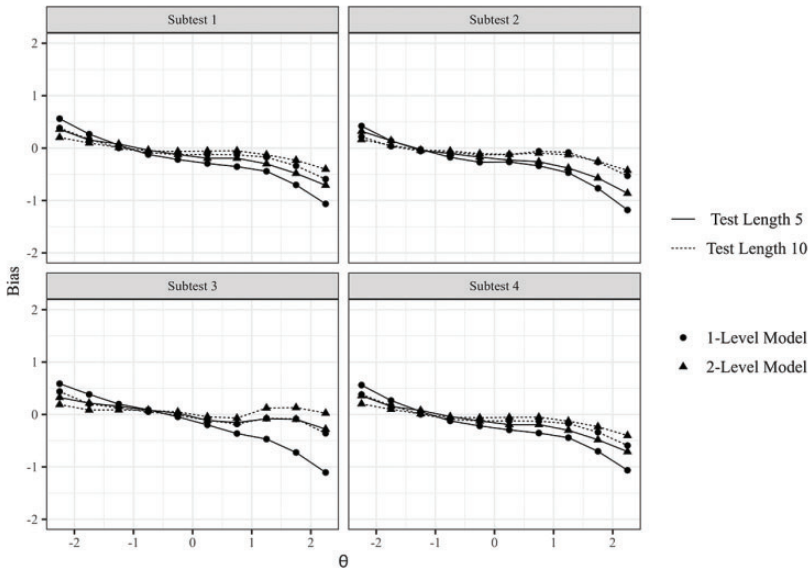


FIGURE 2. Average bias functions for the estimated ability parameters for the four simulated conditions.

Figures 2 and 3, the distribution was actually oversampled to have at least 500 draws for each of the 10 intervals $-2.5(0.5)2.5$. The results are reported for 500 cases randomly selected from each of the intervals for each of the simulated conditions though.

For the condition of two-level adaptive testing, subpool and item selection were entirely according to the Bayesian approach presented in this article. The first subpool was selected averaging the marginal distributions of each θ_h across 10 draws from the posterior samples of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ saved from calibration. For the one-level condition, the subtests were simulated separately with the Gibbs sampler resampling the posterior distributions of the parameters for the selected item

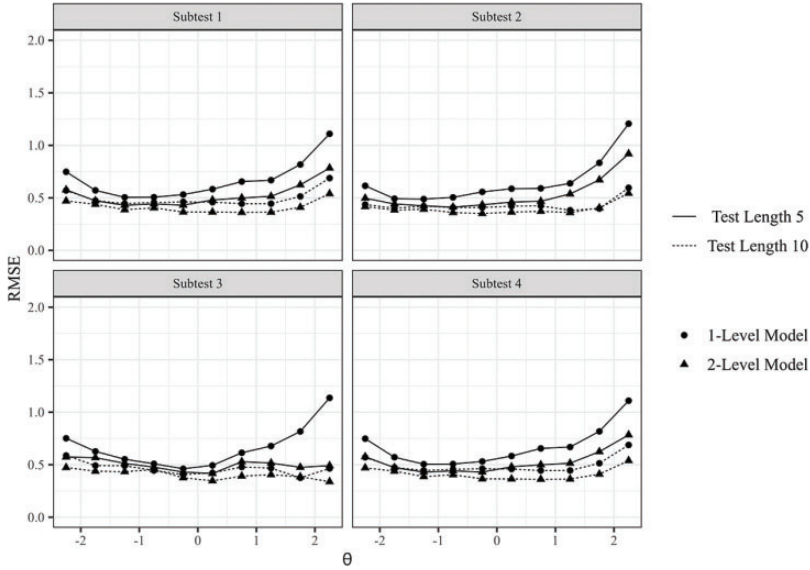


FIGURE 3. Average root mean squared error functions for the estimated ability parameters for the four simulated conditions.

from the current subpool only. Observe that the comparison in this study is thus with a fully Bayesian version of one-level adaptive testing, not the less efficient case of traditional adaptive testing with point estimates of all of its parameters.

As already indicated, the sampler was set to have a burn-in of 5,000 iterations and thinning by a factor of 500. For both conditions, resampling of the parameters was from 500 independent posterior draws saved in the system from the calibration. For security reasons, the authors had access only to the parameter estimates for the items in the pool, not to their content or any of their other attributes. The simulations were therefore run without any content constraints on the selection of the items. At the end of the simulations, the final scores for the simulated test takers were recalculated using the posterior distributions given their complete set of responses in Equation 20.

Results

Figure 2 and 3 show the average bias and RMSE functions for the final scores for each of the four combinations in the simulation. The two-level battery clearly outperformed the one-level version both as for bias and RMSE. Obviously, the same was true for the longer relative to the shorter version of the subtests.

The relatively larger bias and RMSE for the one-level approach at the upper end of the ability scale for the case of four 5-item subtests are due to local

TABLE 2.
Counts of the Paths Through the Two-Level Battery by the Test Takers

Possible Path	Subtests of 5 Items	Subtests of 10 Items
1234	355	326
1243	1,045	1,193
1423	0	1
2314	1,949	131
2341	1,959	2,647
2413	9,637	9,029
2431	1,735	2,673
3124	0	24
3142	0	15
3214	77	13
3241	603	873
3412	0	11
3421	0	104
4213	1,734	1,877
4231	864	1,032
4312	2	36
4321	40	15

scarcity of the items in the pool. As demonstrated by the mean vector in Equation 21, the population had a relatively low ability distribution for each of the subtests, and the item pool matched the population. The 10-item subtests suffered less from the match though.

The most significant result, however, was the performance of the two-level battery with the subtest length of five items relative to the one-level battery with the length of 10. Both the bias and RMSE functions for the two cases were relatively close to each other each, especially at the lower end of the scale. The introduction of the second level of adaptation in the battery thus had the same general effect on the bias and accuracy of the scores as lengthening the subtests by a factor close to two. Alternatively, returning to the dilemma discussed in the Introduction to this article, the results can thus be taken to support the option of increasing the number of subtests in the battery by the same factor without any noticeable loss of quality of scoring while keeping the total testing time constant.

It is also informative to check the different paths through the test battery followed by the test takers. Table 2 shows the counts for each of these paths for the two different subtest lengths collected during the simulation. The first path follows the fixed order, in which the subtests of the battery had been administered during operational testing. However, for both lengths of the subtests, nearly every test takers did profit from the presence of an alternative, more informative path thanks to the second-level of adaptation added to the test battery.

Runtimes

The simulations were run on a computer with a Hexa-core CPU (Intel I7-8700) and 16-GB RAM. The simulation was programmed using *R*, *Version 4.2.2* (R Core Team 2022), with the computationally intensive parts (i.e., Gibbs sampling and item selection) coded in *Rcpp* and *RcppArmadillo* (Eddelbuettel & Sanderson, 2014).

The runtimes for the simulated test takers to update their ability parameters and select the next item ranged from 0.030 to 0.040 second/item. For the selection of the next subtest, the range was 0.214 to 0.243 second/subtest. For the computation of the final scores for each of the subtests according to the procedure discussed directly below Equation 20, the runtimes ranged from 0.821 to 0.892 and 1.096 to 1.170 seconds for the 5-item and 10-item subtests, respectively. The times are small enough for real-world application of the two-level type of adaptive testing proposed in this article.

Discussion

As already hinted at, the combination of a two-level adaptive testing model with the proposed Bayesian algorithm for all parameter updates enables several practical extensions and generalizations of current adaptive testing. One of the options is online calibration of new items. The only thing required is the insertion of an adaptively selected item from a section of field-test items added to the pool toward the end of the subtest. To update the posterior distributions of the field-test parameters, the same Gibbs sampler can be used, but now with an MH step for the parameters of the item and resampling of the current posterior distribution of the test taker's ability parameter. In addition to the advantages of calibration under the actual conditions of testing with optimized sample sizes, the items are immediately available for testing with their parameters directly on the operational scales along with the samples from their final posterior distributions required for the version of adaptive testing proposed in this article. Examples of these options have already been illustrated for traditional one-level adaptive testing by Ren et al. (2017), van der Linden (2018), and van der Linden and Jiang (2020). Another extension is the introduction of response times as a source of collateral information on the test takers' abilities. The Bayesian way of doing so is to replace the second-level ability distribution in the model with the joint distribution of the test takers' ability and speed. For a more traditional approach to the additional use of response times with point estimates for all parameters, see van der Linden (2008). A third option is continued updating of the posterior distributions for the first-level item and second-level parameters for the ability distribution during operational testing. This choice would have both advantages and disadvantages. The greatest advantage would be use of information about these parameters from operational data that so far has been ignored. The disadvantage would be loss of the current posterior independence between the test

taker's ability parameter and the other model parameters given the data, which allowed the Gibbs sampler to just resample the posterior distributions of all model parameters other than the one for the test taker's ability. Further research of the option is necessary to see whether the additional complexity due to loss of posterior independence is worth the effort. Along the same lines, once the fit of a field-test item has shown to be satisfactory, the response to the item can be used to update the test taker's ability parameter as well. The necessity to distinguish between operational and field-test items then disappears completely. The only thing that counts would be the distinction between items with more and less informative priors for their parameters, something a Bayesian approach automatically deals with.

If the assumption of a second-level multivariate normal distribution for the ability parameters in their original metric appears to be untenable and temporary transformation to normality does not work, an alternative is to sample an empirical multivariate distribution of the ability parameter estimates for the population of test takers, for example, a distribution collected during initial use of the traditional one-level version of the test battery. The use of an empirical distribution has the advantage of avoiding any assumption about the shape of the ability distribution for the battery. However, distributions of estimated ability parameters generally have larger variability than an estimate of the distribution of their true values. In addition, extremely large samples of test takers are required to stabilize empirical distributions for larger test batteries, especially the conditional distributions required when the sequence of subtests progresses. More generally, the dilemma faced when choosing between a modeled and an empirical second-level ability distribution is between possible bias in the former and inaccuracy inherent in the latter. However, for the current application, bias as a consequence of a misfitting distribution manifests itself only in the form of a less than optimal order of the subtests for the test takers, *not* as bias in the estimates in their ability parameters. For small test batteries, the alternative may work. But for larger batteries, given the improvement of the estimates already demonstrated in the simulation study above, the authors are therefore in favor of the assumption of a parametric second-level distribution, be it the multivariate normal assumed in the current study or a distribution from any other multivariate family that shows reasonable fit.

Appendix

Shadow-Test Approach

The STA treats adaptive testing as a sequence of full-size tests assembled to be optimal at each new update of the ability parameter while satisfying the complete set of constraints in force for the adaptive test. Each item administered is the best free item in the next shadow test; the rest of the free items is returned to

the pool. As both optimality and constraint satisfaction hold for each of the shadow tests, the same automatically holds for the completed adaptive test.

In the current context, the approach does not only support within-subtest item selection but also adaptive transition from one subpool to the next. The criterion for the transition is the selection of the next subpool as the one with the first shadow test that has the best value of the objective function among all remaining subpools. The criterion automatically avoids the pitfall of running into the necessity to violate any of the constraints during testing.

The approach is possible through the use of mixed integer programming (MIP) for the assembly of the shadow tests. The application of the MIP methodology includes the introduction of binary decision variables for the selection of the items, modeling of the objective function and constraints to be imposed on the selection in terms of these variables, and a call to software with a standard mathematical solver to calculate the solution to the model prior to the selection of the next item. Let x_{i_h} denotes the binary variable for the selection of item i from subpool h , where $x_{i_h} = 1$ represents the decision to select the item and $x_{i_h} = 0$ not to select it. In the current context, the shadow-test model has as objective maximization of the sum of the posterior expected information in Equation 11 across the items. For the constraints, we only specify the general nature of two formally different types of them. Content constraints are typically categorical in the sense that they impose lower or upper bounds on the number of items to be selected from each of a set of content categories. Let V_{c_h} denotes the sets of items in subpool h that belong to category $c_h = 1, 2, \dots, C_h$ and n_{c_h} an upper or lower bound to be imposed on the category. The other type of constraint controls quantitative attributes as word counts, readability indices, expected response times on the items, and so on. Let q_{i_h} denotes the value of an arbitrary quantitative attribute for item i in subpool h and b_{q_h} the upper or lower bound to be imposed on the sum of these attributes across the items in the test.

The core of the shadow-test model for the selection of the k th item in the adaptive test from subpool h is then

$$\text{maximize } (n_h S)^{-1} \sum_{i_h=1}^{I_h} \sum_{s=1}^S I_{i_h}(\theta_h^{(k-1,s)}; \xi_{i_h}^{(s)}) x_{i_h}, \quad (23)$$

subject to

$$\sum_{i_h=1}^{I_h} x_{i_h} = n_h, \quad (24)$$

$$\sum_{i \in S_{k-1}} x_{i_h} = k - 1, \quad (25)$$

$$\sum_{i_h \in V_{c_h}} x_{i_h} \begin{cases} \geq \\ \leq \end{cases} n_{c_h}, \quad c_h = 1, \dots, C_h, \quad (26)$$

$$\sum_{i_h=1}^{I_h} q_{i_h} x_{i_h} \begin{matrix} \geq \\ \leq \end{matrix} b_{q_h}, \quad (27)$$

$$x_{i_h} = 0, \quad i_h = 1, \dots, I_h, \quad (28)$$

where $\begin{matrix} \geq \\ \leq \end{matrix}$ denotes the choice of a (strict) inequality. In addition to the categorical and content constraints in Equations 26 and 27, the constraints in Equations 24 and 25 are necessary to control the length of the test and guarantee the presence of the set of items S_{k-1} already administered in the shadow test when selecting the k th item in the adaptive test. For a more comprehensive introduction to the STA as well as technical details of its implementation, see van der Linden (2005, chap. 9; 2022).

Authors' Note

The authors are greatly indebted to Qi Diao for her contributions to an earlier study leading to the current research.


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) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Luping Niu  <https://orcid.org/0000-0003-3696-1180>

References

- Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries* (Research Report 77-6). Minneapolis: University of Minnesota, Psychometric Methods Program.
- Cheng, Y, Patton, J. M., & Shao, C. (2015). α -stratified computerized adaptive testing in the presence of calibration error. *Educational and Psychological Measurement, 75*(2), 260–283.
- de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT approach. *Applied Psychological Measurement, 33*(8), 620–639.
- Eaton, M. L. (1983). *Multivariate statistics: A vector space approach*. Wiley.
- Eddelbuettel, D., & Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics & Data Analysis, 71*, 1054–1063.
- Gelman, A., Carlin, J. B., Stern, H. A., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC.

- Giallucca, K. A., & Weiss, D. J. (1979). *Efficiency of an adaptive inter-subtest branching strategy in the measurement of classroom achievement* (Research Report 79-6). Minneapolis: University of Minnesota, Psychometric Methods Program.
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing Markov chain Monte Carlo. In W. R. Gilks, S. Richardson, S., & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 1–19). Chapman & Hall.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.
- Moore, T. M., Sandro, A. D., Scott, J. C., Lopez, K. C., Ruparel, k., Njokweni, L. J., Santra, S., Conway, D. S., Port, A. M., D’Errico, L., Rush, S., Wolf, D. H., Calkins, M. E., Gur, R. E., & Gur, R. C. (2023). Construction of a computerized adaptive test (CAT-CCNB) for efficient neurocognitive and clinical psychopathology assessment. *Journal of Neuroscience Methods*, *15*(386), 109795.
- Niu, L., & Choi, S. W. (2022). More efficient fully adaptive testing with a revised proposal distribution. *Behaviormetrika*, *49*, 255–273.
- Patton, J. M., Cheng, Y., Yuan, K.-H., & Diao, Q. (2013). The influence of item calibration error on variable-length computerized adaptive testing. *Applied Psychological Measurement*, *37*(1), 24–40.
- Plummer, M. (2017). *JAGS: Just another Gibbs sampler* (version 4.3.0). [Computer software]. [https://urldefense.com/v3/__http://www.mcmc-jags.sourceforge.__;!!DZ3fjg!-YrXxO7FX_wSsm94PSliQZE9CubdBuse_qhQN7WHPctVHRf3uouCwDgkGOZIR_Y22XM-FxXGU302taeGb78R1JLV\\$.net](https://urldefense.com/v3/__http://www.mcmc-jags.sourceforge.__;!!DZ3fjg!-YrXxO7FX_wSsm94PSliQZE9CubdBuse_qhQN7WHPctVHRf3uouCwDgkGOZIR_Y22XM-FxXGU302taeGb78R1JLV$.net)
- R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. <https://www.R-project.org/>
- Ren, H., Choi, S. W., & van der Linden, W. J. (2020). Bayesian adaptive testing with polytomous items. *Behaviormetrika*, *47*, 427–449.
- Ren, H., van der Linden, W. J., & Diao, Q. (2017). Continuous online item calibration: Parameter recovery and item utilization. *Psychometrika*, *82*(2), 498–522.
- van der Linden, W. J. (2005). *Linear models for optimal test design*. Springer.
- van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5–20.
- van der Linden, W. J. (2010). Sequencing an adaptive test battery. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 103–119). 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). Chapman & Hall/CRC.
- van der Linden, W. J. (2022). Review of the shadow test approach to adaptive testing. *Behaviormetrika*, *49*, 169–190.
- van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education*, *13*(1), 35–53.
- van der Linden, W. J., & Jiang, B. (2020). A shadow-test approach to adaptive item calibration. *Psychometrika*, *85*(2), 301–321.
- van der Linden, W. J. & Ren, H. (2020). A fast and simple algorithm for Bayesian adaptive testing. *Journal of Educational and Behavioral Statistics*, *45*(1), 58–85.

Authors

WIM J. VAN DER LINDEN is Professor Emeritus of Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; wjvdlinden@outlook.com. His interests include test theory, applied statistics, and research methods.

LUPING NIU currently serves as a psychometrician at National Council of State Boards of Nursing, Chicago, IL, 60601; lupingniu@utexas.edu. His affiliation during the research in this article was with the University of Texas at Austin. His research interests involve computerized adaptive testing, item response theory, and multilevel modeling.

SEUNG W. CHOI is a professor in the Department of Educational Psychology at the University of Texas at Austin, Austin, TX 78712; schoi@austin.utexas.edu. His primary research interests are psychometrics, educational assessment, and measurement of health outcomes.

Manuscript received July 8, 2023

Accepted September 17, 2023