Bengs, Daniel; Kröhne, Ulf; Brefeld, Ulf

Simultaneous constrained adaptive item selection for group-based testing

Journal of educational measurement 58 (2021) 2, S. 236-261

DOI: 10.1111/jedm.12285

Journal of Educational Measurement
Summer 2021, Vol. 58, No. 2, pp. 236–261

Simultaneous Constrained Adaptive Item Selection for Group-Based Testing

Daniel Bengs and Ulf Kroehne
DIPF — Leibniz Institute for Research and Information in Education, Frankfurt
Ulf Brefeld
Leuphana University, Lüneburg

By tailoring test forms to the test-taker's proficiency, Computerized Adaptive Testing (CAT) enables substantial increases in testing efficiency over fixed forms testing. When used for formative assessment, the alignment of task difficulty with proficiency increases the chance that teachers can derive useful feedback from assessment data. The application of CAT to formative assessment in the classroom, however, is hindered by the large number of different items used for the whole class; the required familiarization with a large number of test items puts a significant burden on teachers. An improved CAT procedure for group-based testing is presented, which uses simultaneous automated test assembly to impose a limit on the number of items used per group. The proposed linear model for simultaneous adaptive item selection allows for full adaptivity and the accommodation of constraints on test content. The effectiveness of the group-based CAT is demonstrated with real-world items in a simulated adaptive test of 3,000 groups of test-takers, under different assumptions on group composition. Results show that the group-based CAT maintained the efficiency of CAT, while a reduction in the number of used items by one half to two-thirds was achieved, depending on the within-group variance of proficiencies.

Introduction

Formative assessment has attracted considerable attention by educational researchers in the past 20 years, mostly spurred by the highly influential review by Black and Wiliam (1998), who analyzed a broad corpus of literature and put a spotlight on the potential of formative assessment for the improvement of learning. What makes assessment formative is, as has been rightfully noted, not simply the use of specific instruments (Popham, 2008), but a purposeful integration of adequate instrumentation into processes and practice (Bennett, 2011). The specific requirements raised by the intended processes and the embedding into a didactic context hence necessitate the development of customized measurement instruments.

The digital transformation in education opens up new possibilities and potentials for the design of instruments as technology-based assessment progresses from what once were essentially digital versions of paper-based tests to the use of innovative item formats and complex and interactive simulation-based tasks (Bennett, 2015). Technology-based assessment also enables adaptivity and personalization by the means of Computerized Adaptive Testing (CAT; Weiss, 1982). By selecting test items that match the proficiency level of the test-taker, CAT increases

testing efficiency (Segall, 2005), allowing frequent yet reliable assessments of learning states. Moreover, it offers test-takers tasks of adequate challenge, which increases the chance that feedback generated from the results can be valuable (Hattie & Gan, 2011).

The German VERA (Vergleichsarbeiten, roughly: comparative tests) is a national complete survey of math and German language competencies of students of classes 3 and 8 conducted yearly in alternating sets of federal states. The assessment program serves a dual purpose as it aims not only at monitoring and accountability, but also at school development and the improvement of instructional practices (Ditton, 2008; Helmke & Hosenfeld, 2003). Reports provided to teachers include test scores as well as analyses at item level, such as frequently made mistakes and percentages correct, which are presented in relation to the performance of students from schools with similar background variables (Groß Ophoff, Koch, Hosenfeld, & Helmke, 2006; Nachtigall & Kroehne, 2006). Accompanying materials are provided by state authorities to guide teachers in deriving actionable adjustments to their instructional practice from the feedback (e.g., Lankes, Rieger, & Pook, 2015; Pädagogisches Landesinstitut Rheinland-Pfalz, 2018). For instance, the performance of students on a particular item may deviate substantially from that of the reference group or the teacher's own expectations. In that case, teachers are advised to review the item with particular attention to the competencies required to arrive at a correct solution, which may provide insight into the students' strengths and weaknesses and suggest possible starting points for instructional adjustments (Meevissen et al., 2013). Concerning instructional goals, VERA is intended to stimulate the progression toward competency-oriented instruction and to promote a change in the culture of testing and assessment items. In that respect, encouraging teachers to work at the item level also has the purpose of engaging teachers with material that is deemed exemplary for the assessment of competencies and in alignment with federal educational standards (Meevissen et al., 2013).

The VERA assessment is currently delivered as a teacher-administered paper and pencil test, but by federal decision, it is moving toward computer-based assessment (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2012), with some states beginning to discuss CAT. While the paper-based assessment is well aligned with the practice of working at the item level, the move toward computer-based assessment and CAT presents a challenge in this respect. This is because due to adaptivity and the necessity of using large item pools, the total number of items administered to a group of test-takers (e.g., the students in one classroom) is usually very high. As the intended item-level analyses require a substantial degree of familiarity with the items, which are externally developed, the required effort on the side of the teachers needs to be limited in order to secure ongoing acceptance and support of the program. Figure 1 shows the total number of items used per group when administering a standard CAT based on an item pool of 128 items to groups of different sizes.[1] Even for the shortest test lengths, groups of all but the smallest sizes mostly received well over 70 different items. For the group size of 25 and a test length of 25, the median number of items per group is as high as 123, with some groups that have exhausted the item pool. Familiarizing with so many items would impose a significant burden on teachers and would, in fact,

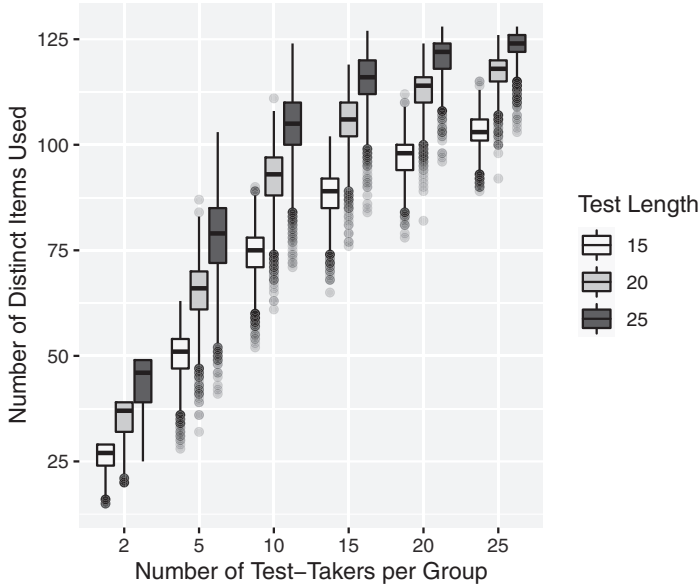

Figure 1. Total number of different items administered per group of test-takers in a simulated assessment using a standard CAT algorithm. Item pool size is 128.

render item-level analyses of results from CAT infeasible. At the same time, very few items would receive a sufficient number of responses to make these analyses meaningful due to the resulting low overlap between the tests of students in each group. Therefore, if assessments like VERA are to leverage the advantages of CAT, effective control of the total number of items per group becomes a requirement.

The purpose of the present paper is to introduce a novel adaptive item selection method that enforces the required limit on the number of different items per group to reconcile adaptive testing with item-level analyses of results in group-based assessments. Our approach makes the number of total items per tested group a parameter of the test specification, which in turn controls the effect of a group-level constraint. The group-level constraint is put into effect by using a test assembly model, which encompasses the tests of a whole group (e.g., school class) of test-takers undergoing assessment concurrently. The assumption of concurrent test-taking made here aligns with usual classroom testing as well as with state testing programs such as VERA. Taking into account the group level in the model for test assembly allows to trade off the range of adaptivity and hence, measurement precision, against the burden on teachers resulting from familiarization with test items. At the same time, item-level adaptivity and full control of aspects of test assembly pertaining to the individual test forms are carried over from standard CAT approaches.

The remainder of the paper is organized as follows. First, we discuss the problem of simultaneous adaptive item selection. We demonstrate how the associated test assembly problem can be treated in the framework of integer linear programming, using two different linearizations of the multiobjective decision problem inherent in

simultaneous optimization of multiple test forms. We then describe the group-based CAT procedure (termed GCAT), which uses the test assembly model, and conduct simulations to evaluate the measurement properties of the GCAT variants. We compare the GCAT method to standard CAT and fixed forms testing and analyze the composition of the assembled tests with respect to overlap of test forms within the groups. We conclude with a discussion of the implications of our results.

Simultaneous Item Selection for Multiple Test-Takers

As laid out above, we assume that a group of test-takers undergoes testing in one session. The test-takers in the group start at the same time and proceed through the test at their own discretion. In this setup, all relevant data collected from the test-takers in the group are available for ability estimation and item selection at any point in time during the test session. This allows to carry out item selection for the individual tests using a model that comprises the test forms of the whole learning group. Aspects of test composition at the group level, in particular, total item usage, can then be controlled by optimizing the individual test forms jointly and simultaneously, subject to constraints at group and individual levels. Updates of the test assembly model are carried out periodically during the test to account for the responses that have been recorded since the last update.

The test assembly model which is at the heart of the GCAT fuses the shadow test approach (STA) proposed by van der Linden and Reese (1998) with models developed for the simultaneous assembly of multiple fixed forms tests (van der Linden & Adema, 1998; van der Linden, 2005) to enable constrained adaptive testing of groups of test-takers. Our model applies the central idea of the STA, which is item selection in terms of complete tests, extending it to the set of test forms selected at the group level. At the individual level, the model contains a shadow test for each test-taker. The purpose of these integrated shadow tests is to ensure that each test-taker‘s test complies with the individual-level test blueprint, which specifies all constraints whose scope is that of a single test, for instance, content balancing. The individual-level shadow tests are coupled by group-level constraints and optimized simultaneously with respect to an objective function that balances the individual tests' information. In the next section, we fix notation and give a precise formal statement of the procedure outlined above.

Test Assembly Model

We will denote by $G = \{1, \ldots, J\}$ the group of test-takers and by $P = \{1, \ldots, I\}$ the item pool, which is assumed to be calibrated under an item response model. By I_i we denote the item information function for item i, which we will assume to be additive such that the test information function for a test $T \subset P$ is given by $I_T = \sum_{i \in T} I_i$.

The GCAT system records all response data submitted by the test-takers. An update of the test assembly model and subsequent item selection may be triggered by either the event of an incoming response or by polling new responses regularly. At the time of the model update, we describe the state of the group of test-takers in a straightforward adaptation of the conventional notation (e.g., van der Linden &

Glas, 2010) as follows. Let L be the test length, which is fixed and uniform across G and denote by $1 \leq k^{(j)} \leq L$ the current position of test-taker j in the adaptive test. Then, the set of items that have previously been administered to j and for which a realization of the response variables has been registered to j is given by

$$S^{(j)}_{k^{(j)}-1} = \left\{ i^{(j)}_1, \ldots, i^{(j)}_{k^{(j)}-1} \right\}. \tag{1}$$

Based on the responses to the items in $S^{(j)}_{k^{(j)}-1}$ an interim ability estimate for test-taker j can be computed (or an initial estimate used if $k^{(j)} = 1$), which is denoted by $\hat{\theta}^{(j)}_{k^{(j)}-1}$. Let $F \subset G$ denote the subset of test-takers who have finished an item by registering a response since the last update of the test assembly model, that is, all $j \in F$ are pending for selection and assignment of $i_{k^{(j)}}$. For the remaining test-takers $j \in G \setminus F$, item $i^{(j)}_{k^{(j)}}$ has been assigned in the course of a previous model update but has not yet received a response and must therefore be held fixed during test assembly.

In order to formulate the test assembly model, it is convenient to introduce decision variables $x_{ij} \in \{0, 1\}$, where $i \in P$ and $j \in G$, such that

$$x_{ij} = \begin{cases} 1, & \text{if item } i \text{ is in the test of test-taker } j \\ 0, & \text{else.} \end{cases} \tag{2}$$

Objective Function. The objective of the GCAT can then be stated as follows. For each test-taker $j \in G$, maximize

$$\sum_{i=1}^{I} I_i(\hat{\theta}^{(j)}_{k^{(j)}-1}) x_{ij} \tag{3}$$

subject to constraints, which will be given in the next section. In general, the constraints at the group level introduce dependencies between the individual test information functions—selecting an item into the set available for the group may increase test information for some test-takers, while decreasing that of others. This conflict of objectives renders the GCAT test assembly model a nontrivial multiple objective decision problem (see Veldkamp, 1999, for a discussion of other multiple objective decision problems in test assembly), which needs to be linearized in order to be solved in the integer linear programming framework. We discuss two different linearization approaches, each of which favors solutions of different character. The first approach combines the objectives for the individual tests into a single, linear objective by forming a weighted sum (Veldkamp, 1999). This yields the objective function

$$f(x) = \sum_{j=1}^{J} w_j \sum_{i=1}^{I} I_i\left(\hat{\theta}^{(j)}_{k^{(j)}-1}\right) x_{ij}, \tag{4}$$

where the weights w_j are positive real numbers. We will refer to this approach as the *weighted* objective function. If used with uniform weights, $w_j = 1$ for all j, the weighted objective function is expected to favor solutions that put more test information around the center (mean) of the latent continuum. This is because under the

usual assumption of a normal distribution of proficiency, extreme realizations of proficiency are rare and thus, the selection of items matching average proficiencies leads to a greater value of the objective than the selection of items that contribute to higher test information for test-takers with more extreme proficiency levels. By using different weightings, for instance, based on prior knowledge about proficiencies, or by changing the weights dynamically during the test, adjustments to this behavior could be realized. Dynamic weighting could be defined in terms of reliability of the interim ability estimates or on the estimates themselves; however, we do not pursue this direction in the present article.

The second linearization approach is the maximization of the minimum of the individual objectives, which was first used for test construction by Boekkooi-Timminga (1989) and for the assembly of multiple test forms by van der Linden and Adema (1998). It can be realized in a linear program by specifying a trivial objective

$$\text{maximize } y, \tag{5}$$

along with the constraint

$$\sum_{i=1}^{I} I_i\left(\hat{\theta}_{k^{(j)}-1}^{(j)}\right) x_{ij} \geq y, \text{ for all } j. \tag{6}$$

We will refer to Equations 5 and 6 as the *maximin* objective function. The maximin objective encodes a different prioritization in the assignment of items than the weighted objective function. Because the test-taker with the least informative test dominates the objective value, a trade-off between the test-takers' tests is favored, which balances test information well between all test-takers in the group.

Individual-Level Constraints. As in the standard STA, a number of technical constraints are required. In the GCAT, they are applied to all test-takers as follows:

$$\sum_{i=1}^{I} x_{ij} = L, \text{ for all } j \text{ (test length)} \tag{7}$$

fixes the test length and

$$\sum_{i \in S_{k^{(j)}-1}^{(j)}} x_{ij} = k^{(j)} - 1, \text{ for all } j \text{ (inclusion of previously administered items)} \tag{8}$$

ensures that the test assembly model takes into account those items that have previously been administered and completed. Similarly, the current item $c^{(j)} = i_{k^{(j)}}^{(j)}$ of the test-takers who have been assigned an item but not responded yet is taken into account by the constraint

$$x_{c^{(j)}j} = 1 \text{ for all } j \in G \setminus F \text{ (inclusion of current item).} \tag{9}$$

The technical constraints laid out so far are, along with the group-level constraints discussed in the next section, sufficient to implement the GCAT bare of any constraints on test content. The flexibility of the integrated STA however allows, in

principle, the application any constraint on individual test content that can be formulated as a linear inequality.

As a practical example, we will discuss the constraints arising from the content balancing, which is applied in the simulation study presented below. Here, the item pool contains items associated with different content areas, $C_1, \ldots, C_F$, which should be represented in each test to a certain amount. By introducing lower bounds b_f and upper bounds B_f for each $f = 1, \ldots, F$, constraints

$$\sum_{i \in C_f} x_{ij} \geq b_f, \quad \text{for all } j \text{ (lower bound for } C_f) \tag{10}$$

$$\sum_{i \in C_f} x_{ij} \leq B_f, \quad \text{for all } j \text{ (upper bound for } C_f) \tag{11}$$

are added.

Group-Level Constraints. The following constraints implement control of the number of different items used at the group level. They are formulated in terms of auxiliary decision variables $z_1, \ldots, z_I \in [0, 1]$. The purpose of z_i is to act as a flag indicating whether item i is selected into the test of at least one test-taker.[2] This amounts to a relation between $(x_{ij})_{i \in P, j \in J}$ and $(z_i)_{i \in P}$, which can be encoded by the set of linear constraints

$$z_i \geq x_{ij}, \text{ for all } j \ (z_i \text{ must be 1 if any } x_{ij} \text{ is 1}) \tag{12}$$

and

$$z_i \leq \sum_{j=1}^{J} x_{ij}, \text{ for all } i \ (z_i \text{ must be 0 if all } x_{ij} \text{ are 0}). \tag{13}$$

The desired bound on the total number of items used for the group then takes the form

$$\sum_{i=1}^{I} z_i \leq M. \tag{14}$$

The value chosen for the upper bound M in Equation 14 as part of the test specification allows to impose a limit on the number of total items. The effects of parameter M on the measurement properties of the GCAT is explored in the simulation study reported below. As in the standard STA (van der Linden & Reese, 1998), if a feasible solution to the initial test assembly problem exists, the subsequent test assembly problems remain feasible for the whole length of the test.

Item Selection. Suppose that $(x_{ij})_{i \in P, j \in G}$ is a solution to the optimization problem defined by Equations 4 to 14. Then, $T_j^* = \{i \in P : x_{ij} = 1\}$ is the shadow test for test-taker j, that is, the optimal constraint compliant test for j at his/her current ability estimate. If $j \in F$, meaning that test-taker j has finished the $k^{(j)}$th item and is pending for item selection, j is assigned the most informative item in T_j^*, which has not been administered to her or him before, which is given by

$$i_{k^{(j)}} = \text{argmax}_{i \in T_j^* \setminus S_{k^{(j)}-1}^{(j)}} I_i\left(\hat{\theta}_{k^{(j)}-1}^{(j)}\right). \tag{15}$$

Table 1
Correlation Matrix of Item Parameters

Variable	b	α	β
b	1.00		
α	−.29	1.00	
β	.14	−.32	1.00

GCAT Procedure

The GCAT proceeds in the following steps:

1. *Initialization*. Set $k^{(j)} = 1$ for all $j \in G$ and set $\hat{\theta}^{(j)}_{k^{(j)}-1}$ to the desired initial value, for example, the population mean or according to prior knowledge. Set $F = \emptyset$ and go to step 2.
2. *Update of test assembly model*. Update the set F to include all test-takers who have submitted a response since the last model update. For all $j \in F$, increment $k^{(j)}$ and update the ability estimate $\hat{\theta}^{(j)}_{k^{(j)}-1}$. Compute the solution $(x_{ij})_{i\in P, j\in J}$ to the test assembly model and carry out item selection.
3. *Termination criterion*. Continue to repeat step 2 periodically until all test-takers have completed L items or an overall time limit is reached.

Simulation Study

We conduct a simulation study to investigate the properties of the proposed GCAT and to compare it with standard CAT and traditional fixed forms testing. The scenario of concurrent test-taking requires simulation of response times as well as response accuracies to account for different speeds at which test-takers may progress through the test. We first discuss the recalibration of the item pool under an appropriate model before we describe the test blueprint regarding content balancing and address data generation and the evaluation criteria used.

Item Pool and Calibration

The item pool used for the simulations was originally assembled and calibrated by Frey, Kroehne, and Born (2011). It is comprised of 128 items designed to measure mathematical proficiency in students of classes 6 to 9 as part of the German Competence Assessment and other large-scale assessments. For the purpose of simulating response times in the present study, the items are recalibrated under a joint model for response times and response accuracies (van der Linden, 2007). Response accuracies are modeled by a 1PL model following the original calibration by Frey et al., who acknowledged the lack of data to fit more complex response models and had selected the presently used 128 items from a set of 155 to ensure satisfactory fit of the remaining items. A log-normal model is used for response times (van der Linden, 2006). The model is fit using the R (R Core Team, 2018) package LNIRT (Fox, Klein Entink & van der Linden, 2007; Fox, Klotzke, & Klein Entink, 2019). Table 1 shows the estimated correlations between item difficulty b, time

discrimination α, and time intensity β. The correlation between person parameters proficiency and speed is estimated at $\rho_{\theta\tau} = -.19$.

Test Blueprint

The content covered by the test items is divided into five areas, $C_1, \ldots, C_5$, associated with five key concepts specified by the curriculum (numbers, measurement, space and shape, functions, data, and chance). According to the test specification, each of these key concepts is required to be represented in each test-taker's test by equal parts. For any test length L divisible by five, the part of the test blueprint pertaining to content balancing is defined by requiring that for each test-taker j and each content area f, the condition

$$\sum_{i \in C_f} x_{ij} = \frac{L}{5}, \text{ (target proportion for } C_f) \tag{16}$$

is satisfied.

Data Set

The base data set for the simulations is generated by drawing a set of $N = 3000 \cdot 25$ test-takers, represented by their proficiency θ and speed τ. The trait vectors are drawn from a bivariate normal distributed population with zero mean, $\sigma_\tau = \sigma_\theta = 1$ and a correlation of $\rho_{\theta\tau} = -.19$ estimated for the calibration sample. The test-takers are randomly assigned to $N_G = 3,000$ groups of each $J = 25$ test-takers. This reflects a simplistic scenario where no systematic effects influence group composition and intraclass correlations of the proficiency and the speed variable are each zero.

In order to arrive at a more realistic setting in which a varying proportion of variance in proficiency is explained by the group level, five derived data sets are created by reducing the variance in θ such that, after adding a group-level effect $U_g \sim N(0, \sigma_U)$, an intraclass correlation coefficient ρ_I of $\{.1, .2, \ldots, .5\}$ results; overall variation is kept fixed at $\sigma_\theta = 1$. This range of ρ_I includes intraclass correlations usually found in educational assessments (e.g., Hedges and Hedberg, 2007a, 2007b). As we are not aware of reports on the typical intraclass correlations for the speed parameter of the response time model, τ is not modified, leaving the intraclass correlation of τ at 0.

Setup of GCAT and Baseline Methods

The GCAT procedure is simulated with both the weighted and the maximin objective function for each of the groups with a limit on the number of items per group of $M = 30, 40,$ and 50 and is labeled GCAT-M. The weighted objective is used with uniform weights, that is, $w_j = 1$ for all j. Test lengths of $L = 15, 20, 25$ are used. As baselines, we compare against a standard CAT procedure (labeled CAT) based on the STA and a fixed forms test optimized by the method presented in van der Linden (2005, pp. 113ff.) using a relative target matching the simulated proficiency distribution in the population. The test assembly model of both the fixed forms and the CAT baseline enforce the same test blueprint as the GCAT. For all methods,

expected a posteriori ability estimation is employed using a standard normal prior and the information criterion is expected Fisher information. Response times are not used as collateral information during ability estimation and solely serve the purpose of simulating the individual pacing of test-takers.

The linear model for test assembly is implemented using the GNU Linear Programming Toolkit (GLPK; Makhorin, 2016) for the solution of the test assembly problem. Newly registered answers are collected once per simulated second and the test assembly model is updated as necessary. The time used to carry out the optimization is not taken into account in the simulation as it is largely implementation dependent and for practical implementations of the procedure, it should be kept as low as possible to avoid substantial waiting times for the test-takers. We do, however, report the computation time recorded by the CAT system and evaluate practical feasibility of simultaneous adaptive test assembly as implemented in our research prototype.

Evaluation Criteria

The results are evaluated based on the following criteria: Measurement precision is evaluated in terms of mean squared error (MSE)

$$\text{MSE} = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\theta}_j - \theta_j\right)^2, \tag{17}$$

and bias,

$$\text{Bias} = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\theta}_j - \theta_j\right). \tag{18}$$

For the computation of conditional MSE and bias, θ values of the data set are binned in the intervals

$$I_k = [-3.5 + k, -3.5 + k + 1), k = 0, \ldots, 6,$$

and averages in Equations 17 and 18 are taken for $\{j : \theta_j \in I_k\}$ and reported at the interval center points.

To provide insight into the effect the limit on the number of items per group has on test composition, we compute the average number of common items between the test forms of two test-takers within a group as follows. Starting from the overlap between a pair of test forms $(S, T) \subset P$,

$$O(S, T) = |S \cap T|, \tag{19}$$

the average O_g is taken for each simulated group g and finally the average across groups is computed by

$$\bar{O} = \frac{1}{N_G}\sum_{g=1}^{N_G} O_g. \tag{20}$$

Conditional mean pairwise overlap is computed by restricting the pairs S, T to test-takers whose proficiency ranges in the interval $I_k, k = 0, \ldots, 6$.

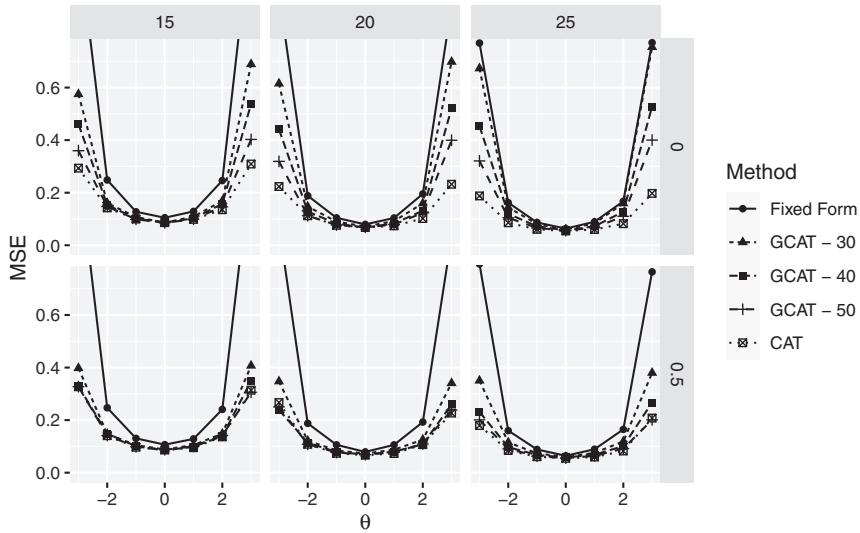

Figure 2. Conditional MSE of fixed forms, GCAT (weighted objective), and CAT, for test lengths 15, 20, 25 (columns), $\rho_I = 0$ (top) and $\rho_I = .5$ (bottom).

Results

Measurement Error and Bias

We investigate measurement properties in terms of conditional and unconditional (average) MSE and bias, comparing the GCAT variants to each other and to fixed forms testing and a standard CAT baseline.

Relation of CAT, GCAT, and Fixed Forms. As should be expected, lowest conditional MSE is attained by the standard CAT procedure, which may utilize the whole item pool in one group (Figures 2 and 3). On the other extreme ranges the fixed forms test, which, limited to a fixed set of items (albeit optimized for the assumed population distribution), is conceded no room for adaptation. The MSE curves of the GCAT interpolate between those of the standard CAT and the fixed forms test, with MSE decreasing with an increasing value for the bound on the number of items per group. In contrast to the fixed forms test and the standard CAT, which are not influenced by group composition, the average MSE realized by the GCAT decreases as intraclass correlation increases (Figure 6). This is expectable because the narrower proficiency spectrum of more homogeneous groups can be covered better with a given number of items from the pool.

The same observations can be made for conditional bias (Figures 4 and 5).

Objective Function. As explained above, the specification of the objective function in the test assembly model for GCAT encodes a preference for the quality of the solution. In accordance with our expectation, differences in the conditional MSE curves between the weighted and the maximin objective are evident in Figures 2 and 3. The weighted objective outperforms the maximin objective in the middle of the

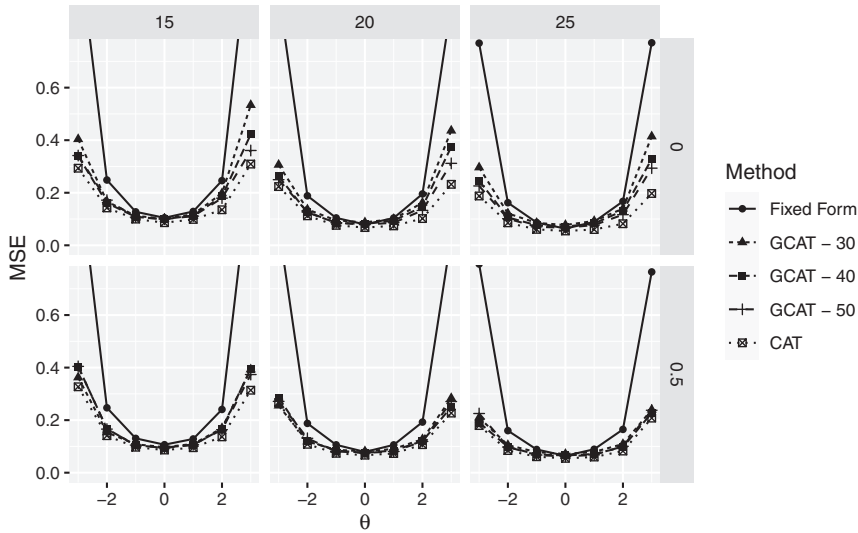

Figure 3. Conditional MSE of fixed forms, GCAT (maximin objective), and CAT, for test lengths 15, 20, 25 (columns), $\rho_I = 0$ (top) and $\rho_I = .5$ (bottom).

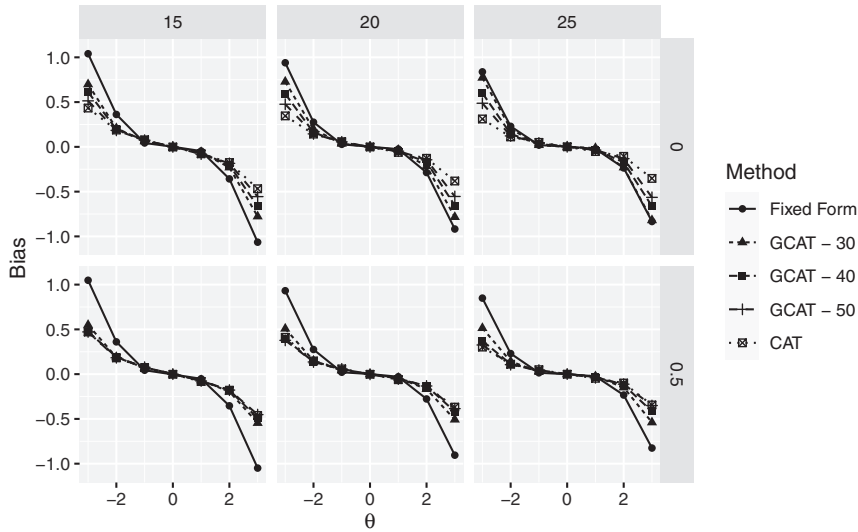

Figure 4. Conditional bias of fixed forms, GCAT (weighted objective), and CAT, for test lengths 15, 20, 25 (columns), $\rho_I = 0$ (top) and $\rho_I = .5$ (bottom).

proficiency distribution, however this comes at the cost of less measurement precision for the more extreme proficiencies. The maximin objective in turn allocates more test information to test-takers exhibiting extreme proficiency levels. Hence the MSE curves for the GCAT using the maximin objective are flatter and follow those of the CATi baseline more closely. Notably, the precision of the maximin GCAT

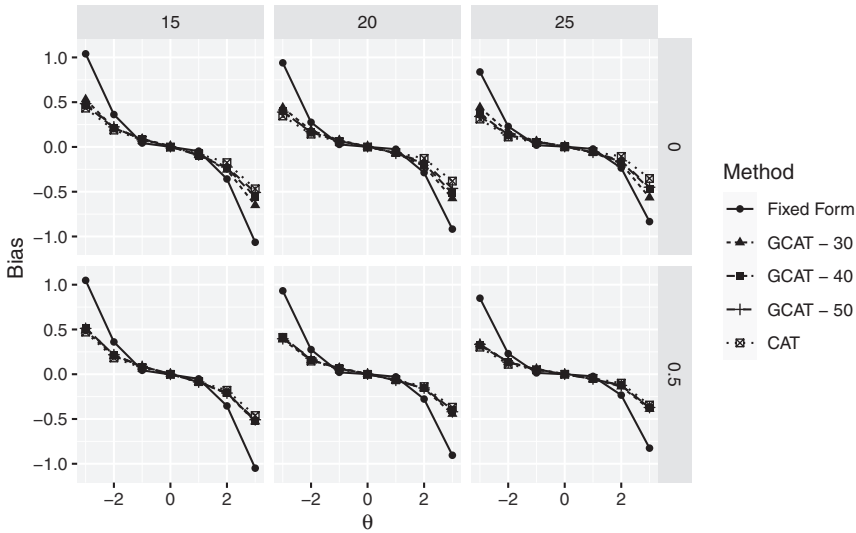

Figure 5. Conditional bias of fixed forms, GCAT (maximin objective), and CAT, for test lengths 15, 20, 25 (columns), $\rho_I = 0$ (top) and $\rho_I = .5$ (bottom).

approximates that of the CAT baseline already for low intraclass correlations when $M = 50$, corresponding to a reduction of the number of items used to 50% of that of CAT, which used more than 100 items on average (cf. Figure 1). When intraclass correlation is higher, $M = 30$ items and hence a reduction to about 30% of the items used per group by the CAT baseline were sufficient to almost attain the same level of precision. With increasing intraclass correlation, the differences between the objective functions decrease and for $\rho = .5$, become very small. The same observations also apply mutatis mutandis to conditional bias (Figures 4 and 5).

Due to the normal distribution of proficiency in the simulated data set, the weighted objective outperforms the maximin objective in terms of average MSE. With increasing intraclass correlation and hence, increasing group homogeneity, the differences between the objectives become less marked and overall MSE decreases (Figure 6).

Test Overlap

It seems reasonable to expect that a reduction of the total number of items per test-taker group should lead to increased overlap in the group members‘ test forms. The simulation results show that indeed, this relationship holds within the GCAT variants (Figures 7 and 8). When comparing the GCAT to standard CAT, the situation is more complex. The standard CAT procedure led to the smallest overlap for average proficiency levels. This can be attributed to the fact that items of average difficulty are relatively abundant in the item pool and could be selected quite freely. Conversely, the high overlap exhibited by standard CAT at the extremes is related to the sparsity of items of high and very low difficulty. The weighted objective GCAT interpolates between the standard CAT and the fixed forms baseline also with respect

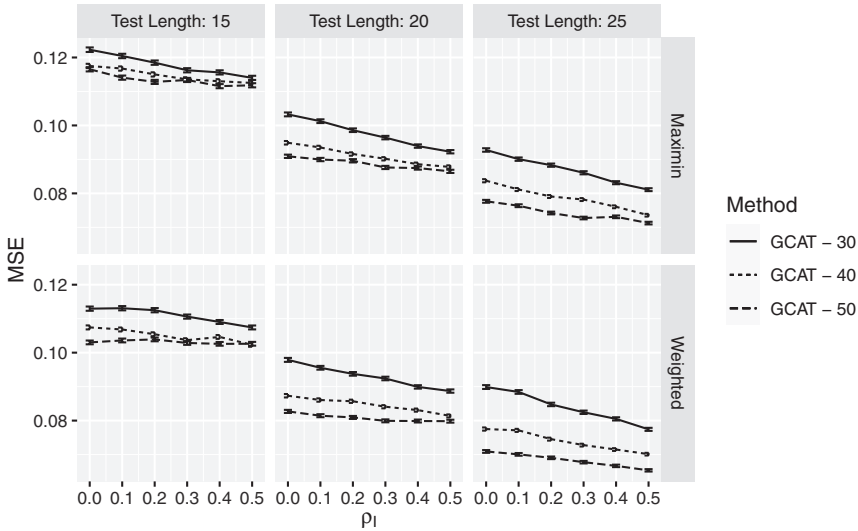

Figure 6. Comparison of weighted and maximin objective for GCAT with varying interclass correlation in terms of average MSE. Errorbars: ± one standard error of the mean.

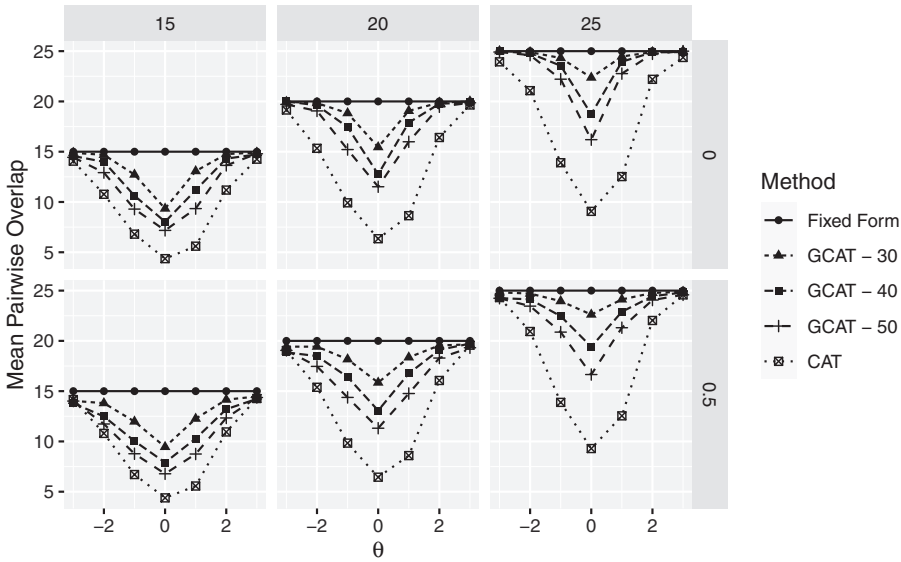

Figure 7. Conditional pairwise overlap between tests, average taken across groups for fixed forms, GCAT (weighted objective), and CAT, for test lengths of 15 (left), 20 (middle), 25 (right), and $\rho_I = 0$ (top), $\rho_I = .5$ (bottom).

to overlap, producing less overlap when the maximum number of items per group is greater (Figure 7). The maximin objective GCAT however exhibits flatter overlap curves with higher overlap than standard CAT in the center and lower overlap in the extremes. This means that overlap between test forms is similar for test-takers of

similar proficiency levels, relatively independent of their actual proficiencies. Overlap decreases for both GCAT variants when intraclass correlation is higher (Figures 7 and 8).

Computation Times

The computational expense needed for optimal test assembly in the experiments presently conducted depends largely on the type of objective function used (Table 2). With the implementation of the test assembly model and the hardware (Intel Xeon E5-2620 CPU equipped with 64GB of RAM) presently used, the median computation time required to optimize the maximin objective is between 2 and 3 seconds, with maximum computation times of over 1 minute (Table 2). The weighted objective is optimized markedly faster, with median computation times just above 1 second and maximal times of about 2.5 seconds. A decrease in computation time can be observed as the bound on the total number of items is increased, coinciding with the intuition that the item selection problem becomes easier when the use of more items is allowed.

Discussion

The GCAT procedure was proposed for testing groups of test-takers under a constraint on the total number of items used per group. The limit on total items per group facilitates formative use of assessment data by teachers, as it reduces the burden resulting from the required familiarization with test items, which often are externally developed. A simulation study was conducted, which confirmed the theoretical expectation that the model of simultaneous test assembly used in the GCAT enabled effective control of the number of items per tested group. The simulations were evaluated with respect to the impact of the limit on total items per group on measurement properties and test overlap, comparing the proposed GCAT to a standard CAT procedure and fixed forms testing. The results indicate that the GCAT maintained the benefits afforded by adaptive testing but effectively limited the number of items used per group. In the condition with no intraclass correlation of proficiencies, the number of items used by standard CAT could be cut roughly by half (to 50 items), without sacrificing much accuracy. In the condition with higher intraclass correlation, about a third of the items used by standard CAT were sufficient to achieve measurement precision virtually at the level of standard CAT. The trade-off between total item usage and measurement precision could be controlled by the parameter chosen for the item limit constraint. As the parameter was varied, measurement precision and bias realized by both GCAT variants interpolated between the extremes represented by fixed forms testing and a standard CAT algorithm. This predictable dependence of measurement properties on the limit on total items per group gives reason to expect that the GCAT can be adjusted to a wide range of requirements arising in assessment practice.

As expected, the constraint on the number of items available per group also influenced test overlap. Here, the choice of objective function led to a qualitatively different outcome. While for both objectives, overlap increased when fewer items were allowed per group, the weighted GCAT produced overlap curves that lay

Table 2
Quantiles of Computation Time for Solving the Test Assembly Problem

Method	Objective	5%	50%	95%	100%
GCAT - 30	Maximin	.83	2.25	15.63	64.92
GCAT - 30	Weighted	.95	1.18	1.9	2.48
GCAT - 40	Maximin	.79	2.18	16.56	67.62
GCAT - 40	Weighted	.9	1.11	1.8	2.42
GCAT - 50	Maximin	.69	1.89	13.51	60.26
GCAT - 50	Weighted	.86	1.05	1.68	2.19

Note. Recorded times include preprocessing, simplex algorithm, and integer optimization.

between those of standard CAT and fixed forms testing, whereas the maximin GCAT exhibited smallest overlap at extreme proficiency levels and more uniform overlap across the latent continuum. The increased test overlap of the GCAT, particularly around the average proficiency levels, however, does not give reason to the usual concerns with respect to test security and overexposure (e.g., Chang & Ansley, 2003; Chen, Ankenmann, & Spray, 2003; Revuelta & Ponsoda, 1998) because of the low-stakes, formative context considered here. In the context of formative assessment as practiced in VERA, it provides a basis for teacher‘s work with item-level results. Two different approaches to the linearization of the multiple objective decision problem inherent in GCAT were considered. The weighted objective function was found to provide better measurement around the center of the proficiency distribution and hence seems the objective function of choice when discrimination between test-takers of medium proficiency levels is desired. As the maximin objective yielded lower MSE and bias for more extreme proficiency levels, it would be the better choice for assessments aimed at detecting individual differences between students in the extremes. However, the magnitude of the described difference between the objective functions depended largely on the homogeneity of the tested groups and decreased with the within-group variance of proficiencies. Hence, if the target population exhibits a high intraclass correlation of proficiencies, the choice of objective function becomes less important.

The improved adaptation exhibited by the GCAT when the intraclass correlation was increased highlights a strength of the CAT approach. Other methods that may be considered for the purpose of limiting the number of items used per class, such as multistage adaptive testing (MST) designs, would require reoptimization based on prior knowledge in order to improve adaptation. Still, MST may represent a middle ground in terms of adaptivity and efficiency and, the requirement of online optimization being absent, makes for a viable alternative that may be easier to implement.

With respect to computational requirements, the GCAT using the maximin objective function posed a greater computational challenge and its implementation based on GLPK might not conform with the real-time requirement imposed by online testing. Optimization of the weighted objective, however, was much faster. Here, computation times measured on our prototypical implementation would be short enough for practical application. Still, in contrast to the small computational effort required

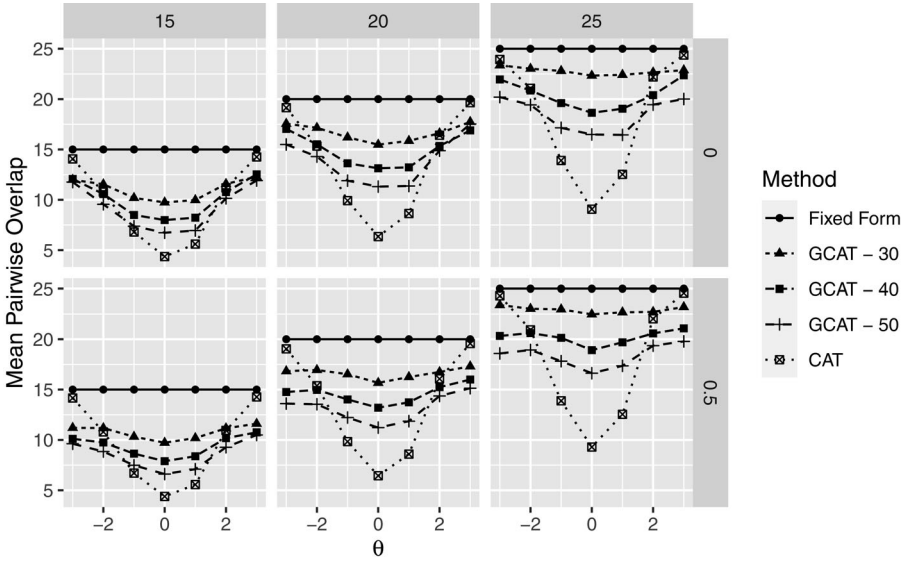

Figure 8. Conditional pairwise overlap between tests, average taken across groups for fixed forms, GCAT (maximin objective), and CAT, for test lengths of 15 (left), 20 (middle), 25 (right) and $\rho_I = 0$ (top), $\rho_I = .5$ (bottom).

when optimizing single test forms, the simultaneous optimization of test forms for a group of test-takers comes at a higher computational expense. For a real-world application of GCAT, especially with the maximin objective, the use of commercial solvers that are known to outperform the open source GLPK by a large margin (Mittelmann, 2019) is advisable. The limitations of the present study include the small size of the item pool that was used in the simulation study and the simple content constraints that were used. While larger item pools would probably exacerbate the motivating problems arising from the high number of different items used per group and low overlap between test forms within the groups, they would also increase the computation times required to solve the test assembly problem. Despite the progress in mixed integer programming solvers and processor speeds, applying the GCAT with very large item pools may be infeasible at present. Yet, the feasibility of GCAT with large item pools and more complex content constraints can only be determined by empirical studies, preferably with optimized implementations and efficient solvers. Still, the timing results from the present simulations show that online adaptive assembly of multiple tests is practically feasible, at least with small to moderate item pool sizes. Integrated into formative assessment programs, GCAT thus can reconcile item-level analyses of testing results with the efficiency afforded by adaptive testing. We hope that by providing this option, GCAT will encourage and facilitate the use of CAT in formative assessment.

Acknowledgement

Open access funding enabled and organized by Projekt DEAL.

Appendix A

The following figures correspond directly to Figures 2, 3, 4, 5, 7, and 8, differing only in that they include the full set of values of intraclass correlation ρ_I used in the simulation study.

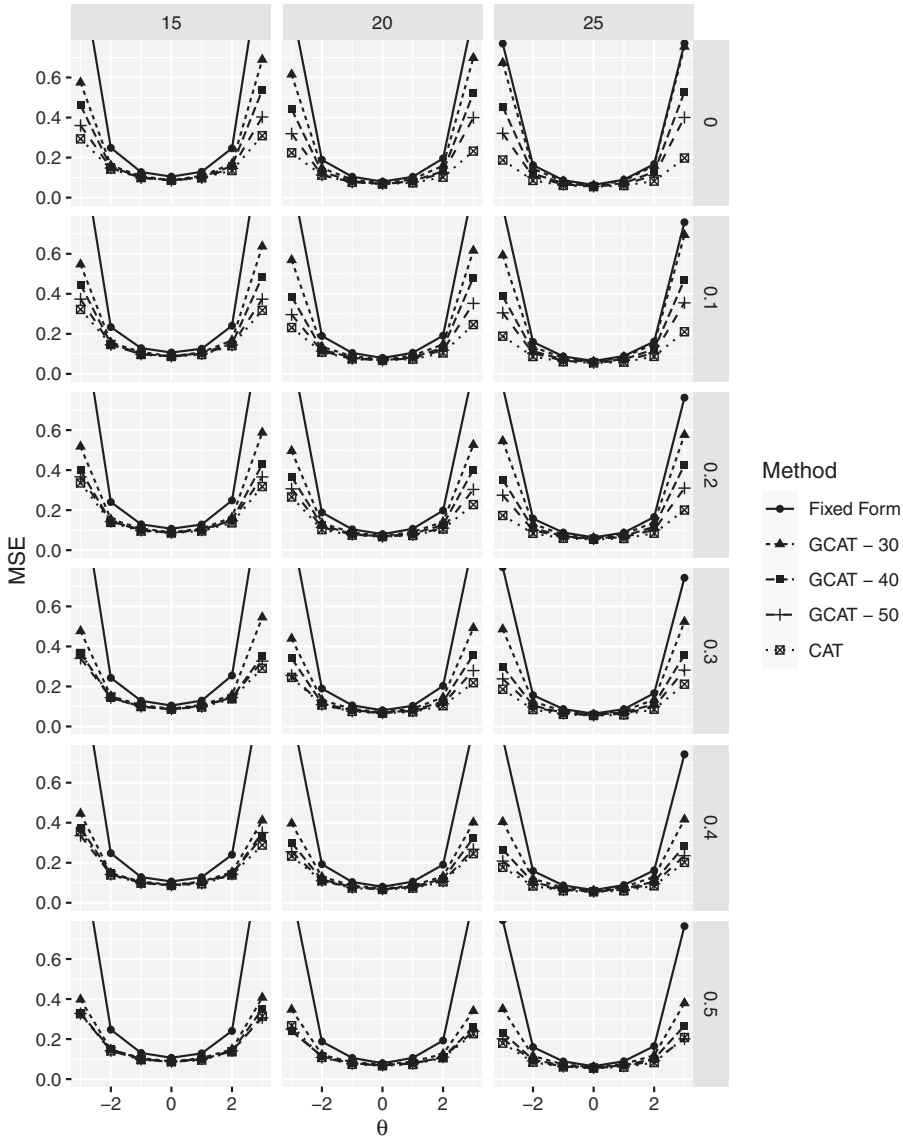

Figure A1. Conditional MSE of fixed forms, GCAT (weighted objective), and CAT, for test lengths 15, 20, 25 (columns), $\rho_I = .0, .1, .2, .3, .4, .5$ (top to bottom).

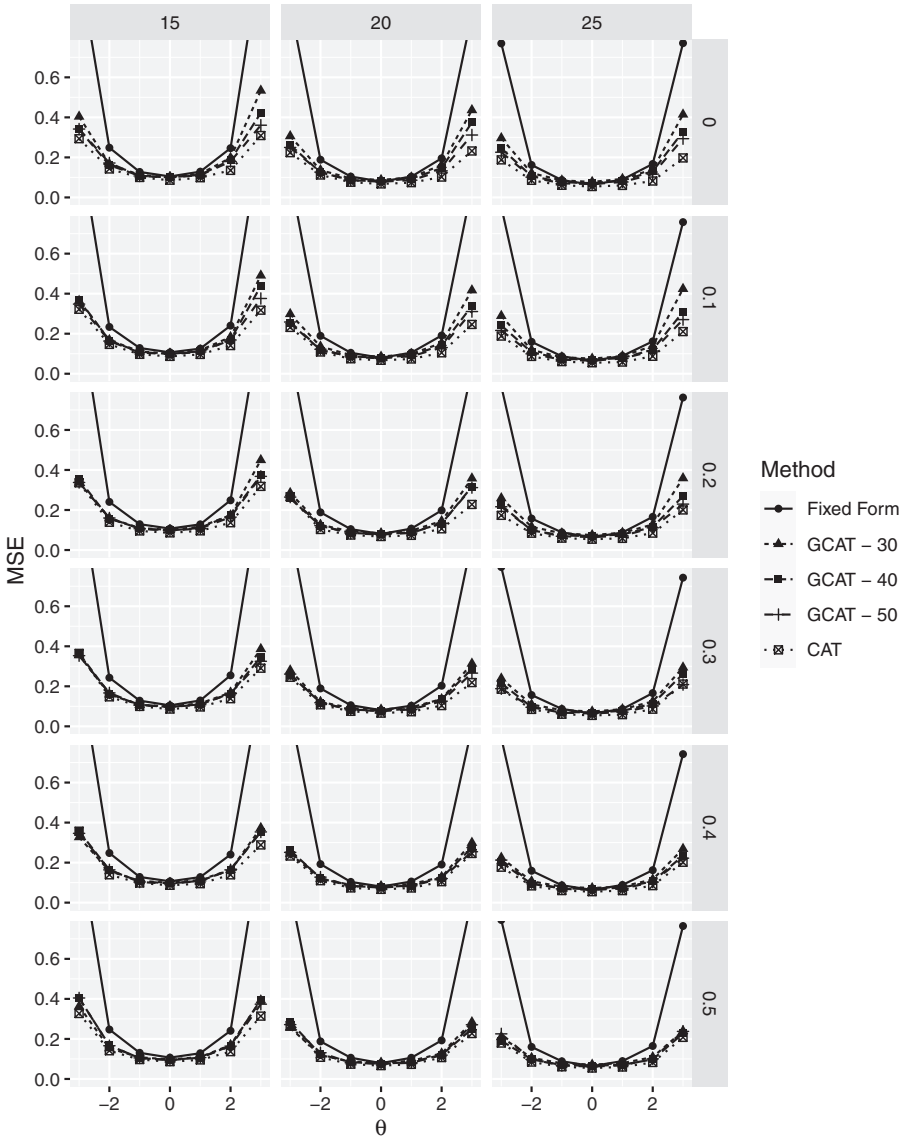

Figure A2. Conditional MSE of fixed forms, GCAT (maximin objective), and CAT, for test lengths 15, 20, 25 (columns), $\rho_I = .0, .1, .2, .3, .4, .5$ (top to bottom).

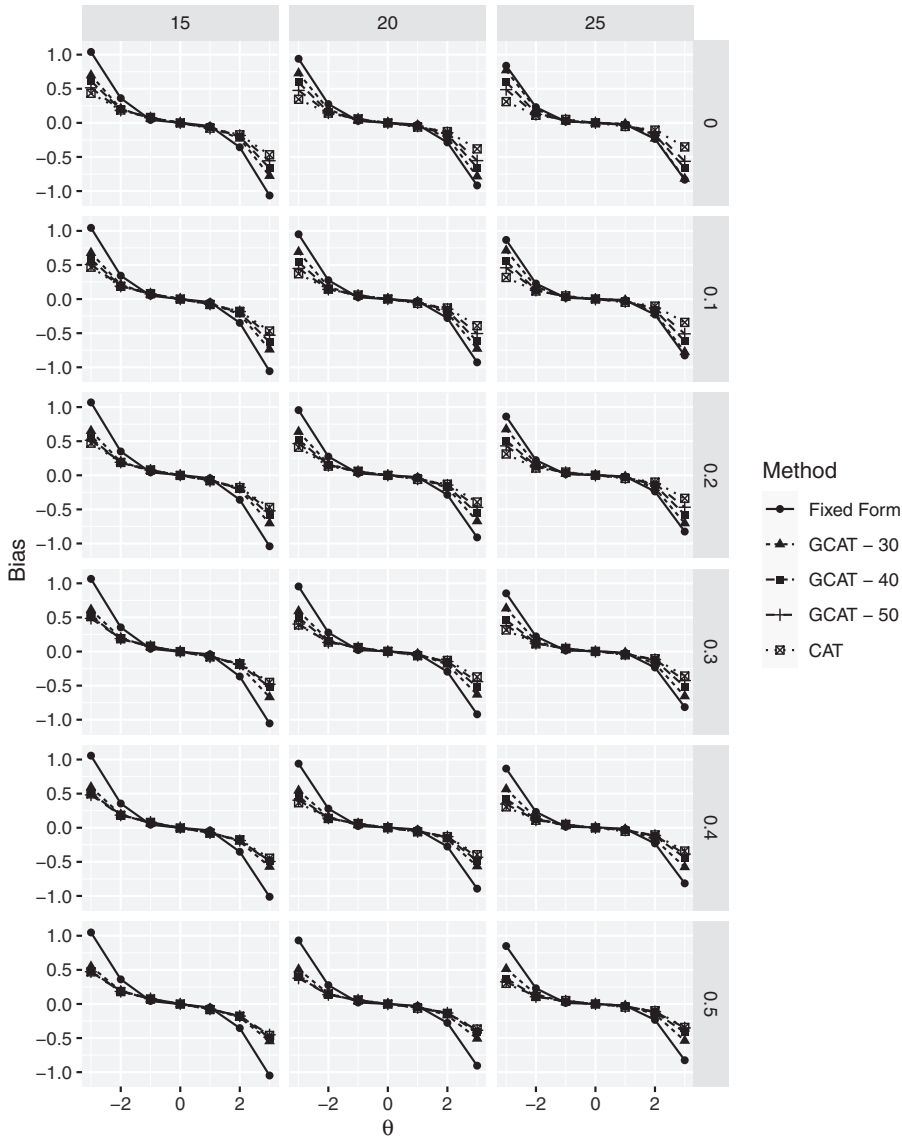

Figure A3. Conditional bias of fixed forms, GCAT (weighted objective), and CAT, for test lengths 15, 20, 25 (columns), $\rho_I = .0, .1, .2, .3, .4, .5$ (top to bottom).

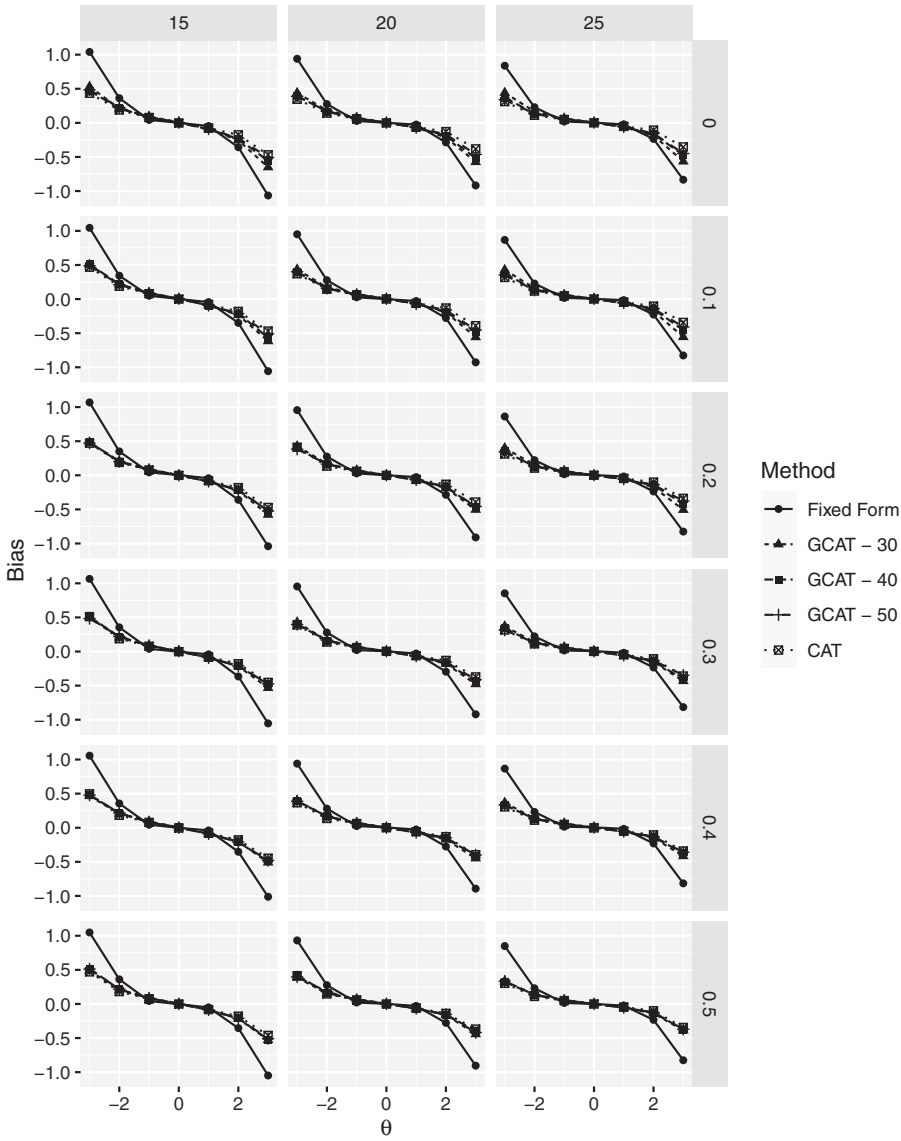

Figure A4. Conditional bias of fixed forms, GCAT (maximin objective), and CAT, for test lengths 15, 20, 25 (columns), $\rho_I = .0, .1, .2, .3, .4, .5$ (top to bottom).

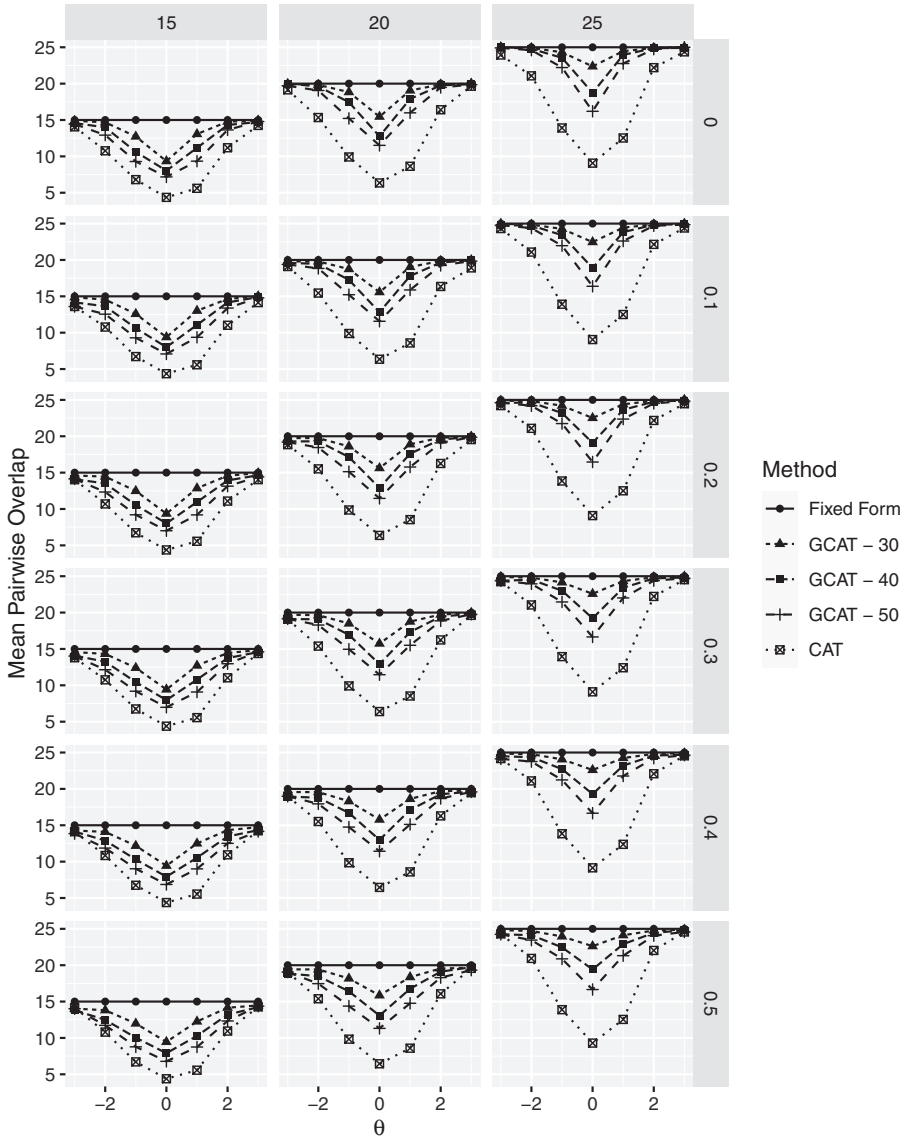

Figure A5. Conditional pairwise overlap between tests, average taken across groups for fixed forms, GCAT (weighted objective), and CAT, for test lengths of 15 (left), 20 (middle), 25 (right) and $\rho_I = .0, .1, .2, .3, .4, .5$ (top to bottom).

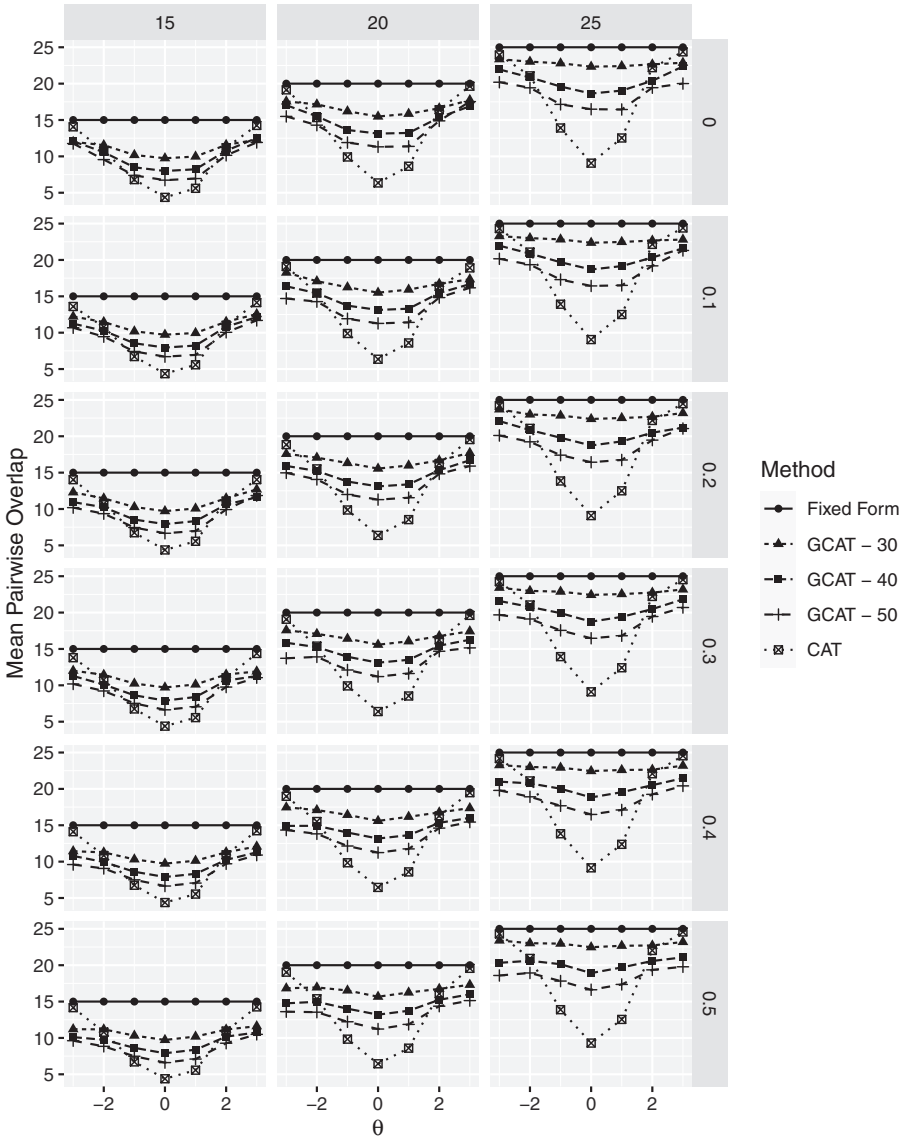

Figure A6. Conditional pairwise overlap between tests, average taken across groups for fixed forms, GCAT (maximin objective), and CAT, for test lengths of 15 (left), 20 (middle), 25 (right) and $\rho_I = .0, .1, .2, .3, .4, .5$ (top to bottom).

Notes

[1]Simulation study; the parameters correspond to those of the CAT baseline described in detail in the section covering the simulation study. Groups were assembled from the population randomly.

[2]z_i can be specified as a continuous variable (which is beneficial for optimization), because as a direct consequence of Equations 12 and 13, it holds that $z_i \in \{0, 1\}$, for all i.

References

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678.

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, *39*(1), 370–407. https://doi.org/10.3102/0091732X14554179.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74.

Boekkooi-Timminga, E. (1989). The construction of parallel tests from IRT-based item banks. *Psychometrika*, *2*(54), 237–247. https://doi.org/10.1080/0969595980050102.

Chang, S.-W., & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *40*(1), 71–103. https://doi.org/10.1111/j.1745-3984.2003.tb01097.x.

Chen, S.-Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, *40*(2), 129–145. https://doi.org/10.1111/j.1745-3984.2003.tb01100.x.

Ditton, H. (2008). Qualitätssicherung in Schulen. *Zeitschrift für Pädagogik*, *53*. Beiheft, 36–58.

Fox, J.-P., Klein Entink, R., & van der Linden, W. (2007). Modeling of responses and response times with the package **cirt**. *Journal of Statistical Software*, *20*(7). https://doi.org/10.18637/jss.v020.i07.

Fox, J.-P., Klotzke, K., & Klein Entink, R. (2019). LNIRT: Lognormal response time item response theory models. Retrieved from https://CRAN.R-project.org/package=LNIRT.

Frey, A., Kroehne, U., & Born, S. (2011). Computerisiertes adaptives Testen im Projekt KomLern (KomLern-CAT). Abschlussbericht, Friedrich-Schiller-Universität Jena.

Groß Ophoff, J., Koch, U., Hosenfeld, I., & Helmke, A. (2006). Ergebnisrückmeldungen und ihre Rezeption im Projekt VERA. In *Rückmeldung und Rezeption von Forschungsergebnissen*.

Hattie, J. & Gan, M. (2011). Instruction based on feedback. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction, Educational psychology handbook series* (pp. 249–271). New York: Routledge. OCLC: ocn548660277.

Hedges, L., & Hedberg, E. C. (2007a). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://doi.org/10.3102/0162373707299706.

Hedges, L., & Hedberg, E. C. (2007b). Intraclass correlations for planning group randomized experiments in rural education. *Journal of Research in Rural Education*, *22*(10), 1–15.

Helmke, A. & Hosenfeld, I. (2003). Vergleichsarbeiten (VERA): eine Standortbestimmung zur Sicherung schulischer Kompetenzen. *SchulVerwaltung Hessen, Rheinland-Pfalz, Saarland*, S. 10–13.

Lankes, E.-M., Rieger, E., & Pook, M. (2015). VERA-3 in Bayern. Retrieved from http://www.isb.bayern.de/download/16050/vera_3_in_bayern.pdf.

Makhorin, A. (2016). GNU linear programming toolkit. Retrieved from http://ftp.gnu.org/gnu/glpk/glpk-4.65.tar.gz.

Meevissen, E., Oldendorf, K., Repschläger, K., Richtering, C., Schröder, M., & Timpter, M. (2013). Handreichung zur Durchführung und Weiterarbeit.

Mittelmann, H. (2019). Benchmarks for optimization software. Retrieved from http://plato.asu.edu/bench.html.

Nachtigall, C. & Kroehne, U. (2006). Methodische Anforderungen an schulische Leistungsmessung - auf dem Weg zu fairen Vergleichen. In *Rückmeldung und Rezeption von Forschungsergebnissen*.

Pädagogisches Landesinstitut Rheinland-Pfalz. (2018). Von Daten zu Taten - Eine Handreichung zum Umgang mit den VERA-Rückmeldungen. Retrieved from https://vera.bildung-rp.de/fileadmin/user_upload/vera.bildungrp.de/E-HR_Vera_Mathe_WEB.pdf

Popham, J. (2008). *Transformative assessment*. Alexandria: ASCD.

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(4), 311–327. https://doi.org/10.1111/j.1745-3984.1998.tb00541.x.

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 429–438). New York: Elsevier.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2012). Vereinbarung zur Weiterentwicklung von VERA.

van der Linden, W. J. (2005). *Linear models of optimal test design*. *Statistics for social and behavioral sciences*. New York, NY: Springer.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

van der Linden, W. J.& Adema, J. J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement*, *35*(3), 185–198.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York: Springer. https://doi.org/10.1007/978-0-387-85461-8.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, *22*(3), 259–270.

Veldkamp, B. P. (1999). Multiple objective test assembly problems. *Journal of Educational Measurement*, *36*(3), 253–266.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473–492. https://doi.org/10.1177/014662168200600408.

Authors

DANIEL BENGS is doctoral researcher in the Centre for Technology-Based Assessment (TBA) at DIPF|Leibniz Institute for Research and Information in Education, Rostocker Straße 6, 60323 Frankfurt am Main, Germany; bengs@dipf.de. His primary research interests include psychometric methods, in particular, computerized adaptive testing.

ULF KROEHNE is postdoc researcher in the Centre for Technology-Based Assessment (TBA) at DIPF|Leibniz Institute for Research and Information in Education, Rostocker Straße 6,

60323 Frankfurt am Main, Germany; kroehne@dipf.de. His primary research interests include psychometric methods and latent variable modeling.

ULF BREFELD is professor for machine learning at Leuphana University, Universitätsallee 1, 21335 Lüneburg, Germany; brefeld@leuphana.de. He is interested in statistical machine learning and structured prediction problems.