Efficiency of Targeted Multistage Calibration Designs under Practical Constraints: A Simulation Study

Stéphanie Berger[12], Angela J. Verschoor[3], Theo J. H. M. Eggen[13], and Urs Moser[2]

[1]University of Twente, [2]University of Zurich, [3]Cito

Author Note

Stéphanie Berger, Research Center for Examination and Certification, University of Twente, Enschede, and Institute for Educational Evaluation, Associated Institute of the University of Zurich; Angela J. Verschoor, Cito, Institute for Educational Measurement, Arnhem; Theo J. H. M. Eggen, University of Twente, Enschede, and Cito, Institute for Educational Measurement, Arnhem; Urs Moser, Institute for Educational Evaluation, Associated Institute of the University of Zurich.

Correspondence regarding this article should be addressed to Stéphanie Berger, Institute for Educational Evaluation, Associated Institute of the University of Zurich, Wilfriedstrasse 15, CH–8032 Zurich.

Email: Stephanie.Berger@ibe.uzh.ch.

Abstract

Calibration of an item bank for computer adaptive testing requires substantial resources. In this study, we investigated whether the efficiency of calibration under the Rasch model could be enhanced by improving the match between item difficulty and student ability. We introduced targeted multistage calibration designs, a design type that considers ability-related background variables and performance for assigning students to suitable items. Furthermore, we investigated whether uncertainty about item difficulty could impair the assembling of efficient designs. The results indicated that targeted multistage calibration designs were more efficient than ordinary targeted designs under optimal conditions. Limited knowledge about item difficulty reduced the efficiency of one of the two investigated targeted multistage calibration designs, whereas targeted designs were more robust.

Keywords: item calibration, test design, Rasch models, simulation

Efficiency of Targeted Multistage Calibration Designs under Practical Constraints: A Simulation Study

Computer adaptive tests (CAT) have become increasingly common in educational assessment owing to substantial technological improvements in computers in recent years. CAT provide a tailored selection of items to a student, thereby measuring his or her ability more efficiently than linear tests (e.g., Lord, 1980; van der Linden & Glas, 2010; Wainer, 2000; Weiss, 1982). Tailored test administration is especially suitable for monitoring the development of student ability over time (Weiss & Kingsbury, 1984; Wright, 1977). In such longitudinal settings, students differ not only in their current ability levels but also in the growth of their ability over time. CAT consider these differences by selecting the most suitable and informative items based on students' performance during test taking.

A calibrated item bank is required to administer CAT. Item calibration involves determining item parameters, such as item difficulty or discrimination before test administration. Such preliminary knowledge about items is essential in CAT to select the most suitable items for a given student during test taking. In the context of CAT, item response theory (IRT) is a common methodological approach to calibrating item banks, as well as for administering CAT (e.g., van der Linden & Glas, 2010; Wainer, 2000). To simplify matters, we limit our study to the Rasch model, an IRT model that includes only one item parameter, namely, item difficulty (Rasch, 1960; Rost, 1996). The Rasch model is used with several operational CAT, such as the National Council Licensure Examination for Registered Nurses (NCLEX-RN; e.g., O'Neill, Marks, & Reynolds, 2005), the Measures of Academic Progress (MAP; Northwest Evaluation Association, 2011) and the STAR Math, Reading, and Early Literacy tests (Renaissance Learning, 2015).

A CAT item bank often consists of several hundred items, particularly if it intends to cover a broad ability range in a longitudinal setting. From a practical perspective, it is very unlikely that a single group of students will respond to all items in the item bank for calibration purposes because doing so would considerably increase the testing time per student (Lord, 1980). Therefore, incomplete calibration designs are very common for calibrating broad item banks from scratch. In such designs, an item bank is divided into several subsets of items, and different subsamples work on these different subsets (Eggen & Verhelst, 2011; Mislevy & Wu, 1996). Incomplete calibration designs reduce testing time per student to a manageable duration. However, such designs require a larger sample than that required by complete calibration designs to achieve the same number of observations per item.

Unfortunately, it is often difficult to find sufficient numbers of students for calibrating items. Thus, there is a need for data collection designs that calibrate items efficiently with a limited number of students. Item calibration under the Rasch model is the most efficient if the ability of the students in a calibration sample matches the difficulty of the items needing calibration (Berger, 1991). Hence, efficient incomplete calibration designs consider students' ability for building subsamples and item difficulty for assembling item subsets. Subsamples with relatively high ability are assigned to relatively more difficult item sets, while subsamples with relatively low ability are assigned to relatively easier item sets (Eggen & Verhelst, 2011).

However, the development of efficient incomplete calibration designs presents two major challenges. On the one hand, it is difficult to build homogenous subsamples for different ability levels. Usually, ability-related background variables, such as grades in school, are used to group students by ability. This is the basic idea of targeted calibration designs (Eggen & Verhelst, 2011; Mislevy & Wu, 1996). Nevertheless, students within such subsamples still vary in their

ability, which results in a loss of efficiency. On the other hand, difficulty of the items is often only vaguely known prior to calibration, which makes it challenging to assemble item sets of similar difficulty and to assign the items to students of corresponding abilities. The lack of knowledge about item difficulty prior to calibration has been considered in a few studies that have investigated the calibration of a small number of field-test items during the administration of operational CAT (e.g., Ali & Chang, 2014; Kingsbury, 2009). Furthermore, Makransky and Glas (2010) have explored different strategies for automatic online calibration of new item pools with unknown item parameters. However, this study focused on ability estimation during the calibration phase and not on item parameter estimation itself (cf. Ali & Chang, 2014). Studies that have compared the efficiency of incomplete calibration designs for calibrating an item pool from scratch have neglected the practical constraint of unknown item parameters by assuming that the difficulty of the items is known (e.g., Berger, 1991; Stocking, 1988).

We address this gap by extending previous research on the efficiency of incomplete calibration designs in two ways: first, we investigate the extent to which it is possible to improve the match between student ability and item difficulty and, therefore, calibration efficiency under the Rasch model, by narrowing subsamples based on students' performance. To this end, we introduce the concept of targeted multistage calibration designs as an extension of targeted calibration designs. Second, we explore how uncertainty about the items' difficulty affects the efficiency of targeted calibration designs, as well as that of targeted multistage calibration designs.

Accuracy and Efficiency in Rasch Model-based Item Calibration

The efficiency of any item calibration design depends largely on the underlying methodological approach. IRT refers to a family of models that express the probability of a student solving an item correctly as a function of student ability and item difficulty (Lord, 1980). For the

Rasch model—the simplest unidimensional IRT model—the probability of a student solving a specific item correctly is given by

$$P(X_{ij} = 1|\theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} = p_{ij}, \quad (1)$$

where θ_i represents the ability of student *i*, β_j is the difficulty of item *j*, and p_{ij} corresponds to the probability that student *i* solves item *j* correctly (Rasch, 1960; Rost, 1996; Strobl, 2012). Within this framework, item calibration refers to establishing model fit and estimating the item difficulty parameter β_j (Eggen & Verhelst, 2011; Vale & Gialluca, 1988). Item difficulty can be estimated by following various maximum likelihood estimation procedures, and the resulting standard error $SE(\hat{\beta}_j)$ serves as a measure of the accuracy of the estimate. The standard error of the estimated difficulty of item *j* can be approximated as (Rost, 1996)

$$SE(\hat{\beta}_j) \sim \sqrt{\frac{1}{\sum_{i=1}^{N} p_{ij}(1 - p_{ij})}}. \quad (2)$$

On the one hand, we can infer from Equation 2 that the calibration sample size is a key factor for calibrating item parameters accurately according to the Rasch model (e.g., Rost, 1996; Wright, 1977). The larger the number of students *N* who provide responses for a specific item, the greater is the accuracy of item difficulty estimation. According to Equation 2, we can conclude that the standard error $SE(\hat{\beta}_j)$ is inversely proportional to the square root of the number of students *N*:

$$SE(\hat{\beta}_j) \sim \frac{1}{\sqrt{N}}. \quad (3)$$

Thus, a reduction of 50% in the standard error $SE(\hat{\beta}_j)$ requires quadrupling the sample size subject to the constraint that the properties of the sample, as well as those of the item pool, remain constant.

On the other hand, we can deduce from Equation 2, that the accuracy of item difficulty estimation depends on the relationship between item difficulty and ability of the students in a calibration sample (Berger, 1991; Eggen & Verhelst, 2011; Stocking, 1988; Wright, 1977). Item difficulty can be estimated most accurately under the Rasch model if the mean ability of the sample is close to the difficulty of the items and if the standard deviation of the students' ability is small (Berger, 1991; Rost, 1996; van der Linden, 1988). If θ is equal to β, we have a probability of 50% for solving item *i* correctly, which in turn, results in minimization of $SE(\hat{\beta}_j)$.

In sum, we can calibrate item difficulty most accurately under the Rasch model if the calibration sample is large and if it includes students with abilities that correspond to the items' difficulty. Enhancing the calibration sample is often difficult in practical settings, which leaves optimization of the relationship between item difficulty and student ability as the only option for improving the accuracy of item difficulty estimation. As shown by Berger (1991), the fit between these two variables depends largely on the calibration design.

Designs for Item Calibration

We distinguish three different types of calibration designs that take item difficulty and student ability into account in optimizing item calibration: (1) targeted calibration designs, (2) multistage calibration designs, and (3) targeted multistage calibration designs, which refer to a combination of the first two designs. We describe all three designs based on an example and discuss the extent to which the designs allow for efficient item calibration under ideal conditions; that is, if the difficulty of all items in the design is known. Practical constraints of limited knowledge about difficulty of the items in an item pool will be discussed in the subsequent section.

Targeted calibration designs. Test designs for targeted testing consist of several test booklets of varying mean difficulty, and students are assigned to the most suitable booklet by means of ability-related background variables, such as grades in school (Eggen & Verhelst, 2011; Mislevy & Wu, 1996). Students can be classified and, therefore, sampled based on such background variables before test administration. The main goal of targeted testing is to estimate student ability more precisely and to prevent students from getting discouraged by items that are not suited to their ability level. However, the same mechanism, that is, assigning students to test booklets based on background variables, can also be used to improve the accuracy of item difficulty estimation during calibration.

Figure 1 shows a basic example of a targeted design on the left. In the example, students are grouped by a background variable y (e.g., grades in school) such that the first subsample with $y = 1$ is assumed to have lower mean ability than the second subsample consisting of students with $y = 2$. This results in two subsamples with greater homogeneity in terms of abilities than the overall sample. Similarly, the item pool is divided into two test booklets such that the first booklet contains mainly items that are easier than those in the second booklet. The overlap between the two booklets (i.e., linking items of intermediate difficulty) ensures that all items can be calibrated on the same scale by using IRT calibration methods (Eggen & Verhelst, 2011; Kolen, 2007; Kolen & Brennan, 2014). Variance in item difficulty is lower within a single booklet than in the total item pool.

According to the example design, students from group $y = 1$ are assigned to the easier test booklet, while students from group $y = 2$ are assigned to the more difficult test booklet. This optimization of the relationship between item difficulty and student ability allows for more accurate

estimation of item difficulty over the entire item pool than random assignment of students to random sets of items from the same item pool. However, this simple example design has one significant disadvantage: linking items that are included in both booklets have twice as many observations as the other items in the design. Consequently, item difficulty of the linking items is estimated much more accurately than that of the remaining items (see Equation 3). This imbalance can be corrected by using more complex designs, as we will show later.

Multistage calibration designs. Multistage tests consist of several parts (i.e., stages), which, in turn, include multiple item sets—called modules—of varying difficulty (Yan, von Davier, & Lewis, 2014; Zenisky, Hambleton, & Luecht, 2010). Students' performance in the first stage determines whether they receive an easier or a more difficult module in the second stage. The decision is based on predefined routing rules. Several studies have shown that multistage tests are more efficient than linear tests in estimating student ability owing to closer alignment between item difficulty and student ability (e.g., Yan et al., 2014). Such an alignment is also advantageous from the viewpoint of estimating item difficulty more precisely (Eggen & Verhelst, 2011; Glas & Geerlings, 2009; Zwitser & Maris, 2015).

To define modules and routing rules that lead to an optimal match between items and students, knowledge is required about ability distribution in the sample, as well as about item difficulty. The routing rules control which and how many students reach a certain module in the second or any subsequent stage. Thus, they determine the ability distributions of the subsamples, as well as the number of observations per item for the modules. For calibration designs, it is desirable to control the number of observations per item such that the difficulty parameters of all items can be estimated with comparable accuracy (Glas & Geerlings, 2009).

Figure 1 shows a basic example of a multistage design on the right. In the first stage, all students start with the same module. This routing module comprises items of intermediate difficulty, and it establishes a link between the two modules in the subsequent stage. Module selection in stage 2 depends on the student's score s in the routing module and on the cut-off score c. Students with a low score (i.e., $s < c$), and thus low ability, are directed to the easy module in stage 2, while students with a high score (i.e., $s \geq c$) are directed to the difficult module. This ability-based routing leads to closer alignment between item difficulty and student ability in the second stage, and—under the condition of an equal number of observations per item—it increases the accuracy of item difficulty estimates.

Again, one complication of this simple example design is that the number of observations per item varies among different groups of items. The items in the first stage are solved by the entire sample, but the sample is divided between the two modules in the second stage. This imbalance can be corrected by using more complex designs with parallel modules (i.e., including two equally difficult modules in stage 1), as we will elaborate later.

Targeted multistage calibration designs. Targeted multistage calibration designs refer to a combination of the two previously presented designs. Figure 2 shows an example of such a design that considers two different subsamples and includes two stages. Again, knowledge about item difficulty is an essential precondition for defining modules characterized by different mean difficulty, as well as for specifying appropriate routing rules. In the first stage, students are either assigned to an easier or a more difficult routing module based on the background variable y (i.e., targeted assignment). Assignment in the second stage depends on performance within the first modules and on the routing rules related to each module. According to the example design in Figure 2, students from group $y = 1$ with a low score in the easy routing module are guided to the

easy module within stage 2. Students from group $y = 1$ with a high score in the easy routing module (i.e., $s_1 \geq c_1$) and students from group $y = 2$ with a low score in the difficult routing module (i.e., $s_2 < c_2$) are both directed to a module of intermediate difficulty, which allows for a link among all modules in the design. Last, but not least, students from group $y = 2$ with a high score in the difficult routing module are directed to the difficult module of stage 2.

Targeted multistage calibration designs have the advantage that they consider preliminary background information, as well as performance information, for optimizing the fit between item difficulty and student ability. In the first stage, background information is used to select the most suitable routing module, which results in an improved alignment between item difficulty and student ability compared to traditional multistage designs. In the second stage, the assignment no longer depends on background variables as in targeted designs, but rather on the performance of the students. Performance is most likely a better predictor of the students' true ability than an ability-related background variable such as grades in school.

Similar to multistage calibration designs, it is important to control the number of observations per item in targeted multistage calibration designs because this factor has a large influence on the accuracy of item difficulty estimation. From this viewpoint, the design in Figure 2 is not ideal for calibration purposes because the number of observations is higher for the items in the two modules of stage 1 (half of the total sample) than for the items in the three modules of stage 2 (one-third of the total sample). As for the other two designs, we will present options for controlling the number of observations per item by developing more complex targeted multistage calibration designs.

Uncertainty of Item Difficulty during Test Construction

When arranging items within a given calibration design, test developers need to know the item difficulties to locate items optimally within the design (Glas & Geerlings, 2009). For example, when implementing a targeted calibration design, as shown in Figure 1, a test developer would need items from three different difficulty levels, namely, easy items for group 1, difficult items for group 2, and intermediate items for linking. Knowledge about item difficulty is even more relevant when constructing modules for multistage designs and for defining the related routing rules. However, usually, no empirical information about item difficulty is available before item calibration. Therefore, the decision about item distribution within the calibration design and the definition of the routing rules depend mainly on the expertise and experience of the test developer and/or other involved experts.

Several studies have investigated the accuracy with which experts, such as test developers, content experts, or item authors, can rate item difficulty, and they have found moderate to high correlations between the ratings and the empirical item difficulties (e.g., Bejar, 1983; Hambleton & Jirka, 2006; Sydorenko, 2011; Wauters, Desmet, & van den Noortgate, 2012). The accuracy of difficulty ratings depends on several factors, such as the content, item type, training of the judges, and number of judges (Hambleton & Jirka, 2006, pp. 407–408).

We conclude from these findings that the distribution of items across modules of different target difficulties might deviate in a practical setting from the optimal distribution in a theoretical setting (i.e., optimal condition), where the difficulty of all items is known. Owing to missing empirical data, test developers might fail to assign all items to the most suitable location within a calibration design. Instead, they might include some easy items in test booklets or modules with

high target difficulty or vice versa, owing to over- or underestimation of item difficulty. This results in more heterogeneous item sets than those under the optimal condition. The number of misplaced items depends on the accuracy with which experts can predict item difficulty.

The Present Study

The aim of this study is to answer two research questions: first, we investigate whether we can achieve higher efficiency in item calibration with targeted multistage calibration designs than with targeted calibration designs. If item difficulty is known during test construction, it is possible to optimize the positions of items within the design and to assemble item sets of homogenous difficulty. For this optimal condition, we assume that a targeted multistage calibration design improves the fit between item difficulty and student ability and, therefore, increases the accuracy of item difficulty estimation. Furthermore, we hypothesize that the efficiency gain of a targeted multistage calibration design depends on the exact design specification (i.e., composition of the different modules and the related routing rules).

Second, we examine how limited a priori knowledge about item difficulty affects the efficiency of both targeted calibration designs and targeted multistage calibration designs. If a priori knowledge about item difficulty is limited, we expect that misplaced items impair the efficiency of all calibration designs. In addition, we hypothesize that targeted multistage calibration designs are impaired to a greater extent by misplaced items than targeted calibration designs because knowledge about item difficulty is needed not only to construct the modules but also to specify the routing rules. Finally, we assume that the degree of vulnerability of targeted multistage designs to misplaced items depends on the exact design specifications.

All hypotheses are addressed in a simulation study in which we vary the calibration design, as well as the accuracy of item distribution, across the different booklets or modules within each design (i.e., number of misplaced items).

Method

To highlight the practical constraints to item calibration, we embedded our simulation study in a practical context, namely, the development of an adaptive online item bank for formative assessment in northwestern Switzerland (Berger, Moser, Verschoor, & Eggen, 2015; Berger, Verschoor, Moser, & Oostlander, 2014). The aim of this online item bank is to assess students' ability and to monitor their progress throughout compulsory school.

Ability Distributions and Item Pool

For each simulation run, two samples of 1,300 simulees were drawn randomly from two normal distributions, $\theta \sim N_1(0,1)$ and $\theta \sim N_2(0.8,1)$, to simulate students from two successive grades in school. The selected sample size represented 10% of the total student population from two successive school years in northwestern Switzerland (Berger, Moser et al., 2015), which refers to the expected response rate for a calibration study. The performance progress between the two grades in school was modeled on selected results from a longitudinal study that investigated, among other variables, the performance progress of Swiss students during primary school (i.e., progress in mathematics from grades 3 to 6; Angelone, Keller, & Moser, 2013, p. 35).

We generated an artificial item pool of 180 dichotomous Rasch items with equally spaced difficulty parameters β_j ranging from –1.5 to 2.3 to provide items that cover the range of +/–1.5 standard deviations from the means of both samples. Thus, the ability of approximately 16% of the simulees from the two samples lay outside of this range, which ensured variation in the response patterns for all items. We ensured uniform distribution of item difficulty to investigate the

efficiency of calibration in relation to item difficulty. The size of the item pool was adjusted to the sample size to achieve a reasonable number of observations per item (i.e., approximately 400 observations per item; cf. Wright, 1977, p. 106). Furthermore, we selected an item pool size that allowed us to construct balanced designs for all investigated design types, given a practically relevant test length (see next paragraph for further details). The same item pool was used for all conditions within the simulation study.

Test Designs

The simulation study included three different calibration designs (i.e., a targeted calibration design and two variations of a targeted multistage calibration design) with a multi-matrix structure, which is often used in large calibration studies, as well as a random condition as a baseline. In all four design conditions, the item pool described above was used, and test length was fixed to 30 items, which refers to the number of items that we expect students to solve within one school lesson (i.e., 45 minutes). All designs aimed to achieve an equal number of observations per item. Keeping this factor constant allowed us to focus on the effect of the match between item difficulty and student ability on the efficiency of the different designs.

Targeted calibration design. In the targeted calibration design, the 180 items were divided into six easy and six difficult booklets with 30 items each (see Figure 3). To construct equally difficult booklets for each difficulty level, the items were sorted by difficulty and split into three equally large categories: an “easy” category (items 1–60), “intermediate” category (items 61–120), and “difficult” category (items 121–180). Each category was further divided into six equally difficult modules of 10 items, which resulted in a total of 18 modules (i.e., six modules per category). All modules were included in two different booklets for linking purposes. The six easy booklets contained two easy and one intermediate module each, while the six difficult

booklets included one intermediate and two difficult modules. Each simulee from the low ability group was assigned randomly to one of the six easy booklets, and each simulee from the high ability group was assigned randomly to one of the six difficult booklets.

Targeted multistage calibration designs. We investigated two different targeted multistage calibration designs in our simulation study to analyze the effect of different design specifications on calibration efficiency. Both designs were specified as simply as possible given the item pool size of 180 items, test length of 30 items, and requirement of an equal number of observations per item. The designs consisted of two stages, and each stage included half of the total item pool (i.e., 90 items). In general, items with more extreme difficulty were included in the second stage, where simulees were assigned to the different modules based on their performance in the first stage. Within each stage, the 90 items were divided into six modules of 15 items, resulting in 12 modules in total. The modules were not linked, and each item was included in one module only. Instead, links between the different modules were established by overlapping paths (i.e., various combinations of different modules from both stages).

The two designs differed in terms of module composition and routing. Figure 4 shows the targeted multistage calibration design A (TMST A). In this design, we grouped the 180 items into five difficulty categories. The easiest, intermediate, and most difficult categories were assigned to the second stage and included 30 items each: items 1–30, items 76–105, and items 151–180, respectively. The relatively easy and relatively difficult categories were assigned to the first stage and included 45 items each: items 31–75 and items 106–150, respectively. Within each category, we further divided the items into equally difficult modules of 15 items. In the first stage, simulees from the low ability group were assigned randomly to one of the three relatively easy modules, and simulees from the high ability group were assigned randomly to one of the

three relatively difficult modules. The score in the first stage determined module selection in the second stage. Simulees from the low ability group could reach either an easy or an intermediate module (c_1 = 11); simulees from the high ability group were guided to either an intermediate or a difficult module (c_2 = 5). The two cut-off scores were determined in preceding simulations to achieve an equal number of observations per module in the second stage. The aim was to guide one-third of each sample to the intermediate modules and to guide the remaining two-thirds of each sample to the easy or the difficult modules. The design was linked in two ways: first, simulees from both samples could reach the intermediate modules. Second, the random assignment of students to modules of equal difficulty resulted in multiple combinations of the modules over the two stages (i.e., overlapping paths).

In the second targeted multistage design (TMST B, see Figure 5), we distinguished four instead of five different difficulty levels and split the item pool into four categories of 45 items: items 1–45, items 46–90, items 91–135, and items 136–180. Each category was further divided into three equally difficult modules of 15 items. Simulees from the low ability group started with one of the three relatively easy modules, and simulees from the high ability group were assigned randomly to one of the three relatively difficult modules. Simulees' scores in the first stage determined the selection of difficulty level in the second stage. In both samples, the simulees were directed to one of the three easy modules if they scored lower than the sample-specific cut-off score (c_1 = 11; c_2 = 5). Simulees with equal or higher scores were directed to one of the three difficult modules. Again, the two cut-off scores were determined by preceding simulations to achieve equal numbers of observations per module in the second stage. The aim was to guide three-fourths of the low ability group to the easy modules and one-fourth to the difficult mod-

ules. For the high ability group, the goal was to route one-fourth of the simulees to the easy modules and three-fourths to the difficult modules. Similar to TMST A, the different modules in TMST B were linked in two ways: first, all modules in the second stage could be reached by both samples, and, second, the random assignment of students to modules of equal difficulty resulted in multiple combinations of the modules over the two stages (i.e., overlapping paths).

Random test. We included a random test as a baseline in our simulation study. In this condition, each simulee was assigned to a random selection of 30 items out of the entire item pool. Hence, neither the ability of the target sample nor its performance during test taking was considered during item selection.

Simulation with Limited Knowledge about Item Difficulty during Test Development

To simulate the case with limited knowledge about item difficulty during test development, we manipulated the order of the items in relation to item difficulty and, therefore, the distribution of the items across the different modules in the designs. In the optimal condition with complete knowledge about item difficulty, the items were ordered by difficulty, as described earlier, which resulted in a correlation of $r = 1.0$ between item order and item difficulty. In addition to this full knowledge condition, we investigated two additional conditions in our simulation study. In the first condition, item order and item difficulty were correlated at $r = 0.4$ to simulate a low amount of knowledge during test construction following results reported by Berger, Oostlander, Verschoor, Eggen, and Moser (2015). This condition resulted in a high number of misplaced items. In the second condition, item order and item difficulty were correlated higher ($r = 0.6$) to represent a condition with a medium amount of knowledge during test construction and a resulting medium number of misplaced items. Both conditions were applied to the targeted calibration design, as well as to the two targeted multistage calibration designs, meaning that the

manipulated item order was used as a basis to distribute the items within the designs. Owing to the manipulated item order, item categories were no longer homogenous, but included a few items that exceeded or fell below the envisaged difficulty range. This, in turn, affected the difficulty range and the mean difficulty of the booklets and modules.

Item Response Generation and Calibration

Altogether, the simulation study included 10 different conditions: the three calibration designs (i.e., targeted calibration design and two versions of targeted multistage calibration designs) were combined with the three variations of item distribution (i.e., full, medium, and low knowledge) and completed by the random condition that served as a baseline. For each condition, we generated 20,000 datasets according to the Rasch model. For each dataset, we drew new samples from the two ability distributions described and generated new response patterns consisting of 30 × 2,600 responses. The Multidimensional Item Response Theory (MIRT) software application (Glas, 2010) was used to generate marginal maximum likelihood (MML) estimates of the item parameters from the Rasch model. All designs were estimated using two marginal proficiency distributions, that is, one marginal distribution per sample (cf. Eggen & Verhelst, 2011).

Evaluation Criteria

Distribution of item difficulty per booklet or module. As a descriptive evaluation criterion, we explored the distribution of item difficulty within and among the different booklets and modules. This distribution was predefined under the full knowledge condition, where the item difficulty of each item was known. For the conditions with limited knowledge, the distribution of item difficulty depended on manipulation of the item order.

Bias and root mean square error. To determine the accuracy of item parameter estimation in the different calibration designs, the bias and the root mean square error (RMSE) of each item in each condition were computed. That are,

$$\text{Bias}(\hat{\beta}_j) = \frac{\sum_{k=1}^{20000}(\hat{\beta}_{kj} - \beta_j)}{20000} \quad \text{and} \tag{4}$$

$$\text{RMSE}(\hat{\beta}_j) = \sqrt{\frac{\sum_{k=1}^{20000}(\hat{\beta}_{kj} - \beta_j)^2}{20000}}, \tag{5}$$

where *k* and *j* represent replications and items, respectively, and $\hat{\beta}$ denotes the estimate of item difficulty β.

Mean number of observations per item. As an additional criterion for the efficiency of the different calibration conditions, we investigated the mean number of observations per item over 20,000 simulation runs within each simulation condition. That is,

$$N_j = \sqrt{\frac{\sum_{k=1}^{20\,000} N_{kj}}{20000}}, \tag{6}$$

where *k* and *j* represent replications and items, respectively, and *N* denotes the number of observations. As described earlier, the higher the number of observations per item, the higher is the accuracy of item difficulty estimation. The intention in our setting was to asymptotically achieve equal mean numbers of observations for all items (i.e., $SD_N = 0$) to ensure that the influence of this factor on calibration efficiency was constant for all items.

Results

Distribution of Item Difficulty per Booklet or Module

Figure 6 shows the distribution of item difficulty within the different booklets and modules for each design and knowledge condition by means of boxplots. For the full knowledge condition (dark gray boxplots), this distribution was given by the design specifications. Independent

of the design, the booklets and modules representing different difficulty levels differed clearly in terms of difficulty range and median (e.g., B06 vs. B07). Meanwhile, the booklets and modules that represented one difficulty level had comparable item difficulty distributions (e.g., B01 vs. B02). The range of item difficulty within a single booklet or module was larger in the targeted design than in the two targeted multistage designs. This result is directly related to the number of difficulty categories defined for each design (i.e., two difficulty categories for the targeted design, five for TMST A, and four for TMST B).

By contrast, the simulation with limited knowledge about item difficulty during test construction (i.e., medium and low knowledge) resulted in more heterogeneous booklets and modules (see medium and light gray boxplots in Figure 6). The range of item difficulty within the booklets and modules increased under these conditions, and the medians shifted towards the mean difficulty of the total item pool ($M = 0.400$). Consequently, differences between booklets and modules intended to be equally difficult increased, and differentiation between booklets and modules intended to represent different difficulty levels decreased. These effects were generally more prominent under the low knowledge condition than under the medium knowledge condition. From these findings, we conclude that our manipulation led to the envisaged misplacement of items in the different designs.

Bias($\hat{\beta}$), Mean RMSE($\hat{\beta}$) and Mean Number of Observations per Simulation Condition

This section provides a general overview of the efficiency of the different simulation conditions. Table 1 presents the range of Bias($\hat{\beta}$), the mean of RMSE($\hat{\beta}$) and the related standard error based on the first 1,000 simulation runs under each condition as an indicator of overall efficiency. As indicators of the relative efficiency of the different designs, on the one hand, Table 1

reports the relative decrease or increase in RMSE($\hat{\beta}$) compared to the random condition and targeted design within the corresponding knowledge condition; on the other hand, the table includes for each condition the differences in the number of students needed to achieve the same overall efficiency with the random condition and targeted design (absolute number and percentage of total sample, i.e., 2,600) given Equation 3. This information serves as an indicator of the practical relevance of the reported differences. Also, Table 1 lists the standard deviation of the mean number of observations under each simulation condition, as well as the mean number of observations per difficulty category (i.e., design specific categorization of the modules' difficulty levels).

In all conditions, we observed very little bias and the highest absolute value of Bias($\hat{\beta}$) was found in the Random condition (i.e., Bias($\hat{\beta}$) = 0.012). Independent of the design, the mean RMSE($\hat{\beta}$) was generally lower under the full knowledge condition than under the random or the limited knowledge conditions. The lowest mean RMSE($\hat{\beta}$) was found for TMST B under the full knowledge condition (M = 0.110). The efficiency gain of TMST B was 7% compared to the random condition and 4% compared to the targeted design. To achieve the same efficiency with the baseline designs, we would need to increase sample size by 336 and 200 students, respectively. The overall efficiency gain of TMST A under the full knowledge condition was slightly lower (M = 0.112), yet the design outperformed the targeted design (M = 0.115), which, in turn, yielded lower mean RMSE($\hat{\beta}$) than the random condition (M = 0.119). All differences were statistically significant, as can be derived from the small standard errors of RMSE($\hat{\beta}$) in Table 1.

Under the medium knowledge condition, the efficiency gain of TMST B (M = 0.116) and the targeted design (M = 117) were considerably smaller than that under the full knowledge condition. Nevertheless, the gain of TMST B was still statistically significant and practically relevant in terms of the number of students: 114 (4%) additional students would be required to

achieve the same efficiency with the random condition, and 52 (2%) additional students would be required for the targeted design. The mean RMSE($\hat{\beta}$) of TMST A was even higher than that of the two related baselines (M = 0.120), thereby indicating a significant efficiency loss. Similar but more prominent results were found for the low knowledge condition: TMST B (M = 0.118) was slightly more efficient than the two baseline conditions, which showed a comparable mean RMSE($\hat{\beta}$) (M = 0.119), whereas TMST A resulted in a considerable efficiency loss (i.e., –6%; M = 0.125). To compensate for this loss, we would need to increase the sample size by more than 300 students or 12% of the total sample.

The standard deviations of the mean number of observations displayed in column 7 of Table 1 provide some insights into the roots of efficiency differences among the different conditions. The mean numbers of observations were well balanced in the random and the targeted designs for all three knowledge conditions. On average, all items were assigned to approximately 433 simulees, which corresponded to one-sixth of the total sample (i.e., 2,600 simulees). Moreover, variations in TMST B were very small under all simulation conditions (SD_{full} = 0.114; SD_{Medium} = 6.315; SD_{Low} = 2.925), so only the modules in stage 2 differed slightly in their mean number of observations per item (i.e., M_1 and M_4). However, for TMST A, the number of observations per item depended largely on the available knowledge about item difficulty. Low knowledge was associated with a considerable increase in the standard deviation of the mean number of observations per item (SD = 90.950) compared to the medium (SD = 60.958) and the full knowledge conditions (SD = 0.114). Although the modules of stage 1 showed the expected mean numbers of observations ($M_2 = M_4$ = 433), the mean number of observations varied considerably among modules of different difficulty levels in stage 2 (i.e., M_1, M_3, and M_5).

In sum, these general findings were consistent with our hypothesis that targeted multistage calibration designs calibrate item difficulty more accurately than targeted designs if item difficulty is known. Furthermore, the results support our hypothesis that limited a priori knowledge about item difficulty impairs the efficiency of all calibration designs. In line with our expectations, the results suggest that limited knowledge does not affect all designs to the same extent. However, counter to our hypothesis, one of the two targeted multistage designs outperformed the targeted design under the limited knowledge conditions.

RMSE($\hat{\beta}$) and Mean Number of Observations per Item

Optimal condition with full knowledge. Figure 7 shows RMSE($\hat{\beta}$) for each item in relation to its true difficulty β for all four designs under the full knowledge condition. For the random baseline condition, RMSE($\hat{\beta}$) was the lowest for items with a difficulty of $\beta \approx 0.400$, which corresponded to the mean difficulty of the item pool, as well as to the mean ability of the entire sample ($M_\beta = M_\theta = 0.400$). For items with more extreme difficulty, RMSE($\hat{\beta}$) increased, which resulted in a U-shaped relationship between RMSE($\hat{\beta}$) and β over the entire item pool. Given the normal distribution of ability in the two samples, the number of simulees with abilities that matched the difficulties of the extreme items was limited. Therefore, the difficulty of these items was estimated less precisely than that of the items close to the mean abilities of the two samples ($M_{\theta_1} = 0.000$; $M_{\theta_2} = 0.800$).

The curves related to the other three designs showed a similar tendency to a U-shaped course, but they were characterized by a few kinks (i.e., rapid increase or decrease in RMSE($\hat{\beta}$) between two items). These kinks divided the curves into multiple sections, which corresponded to the difficulty categories that were part of the design specifications. The gray curve representing the targeted design included two kinks that divided the curve into three sections. The first

section included all items from the "easy" category that were only administered to sample 1, the second section corresponded to the linking items ("intermediate" difficulty category), and the third section represented the items from the "difficult" category that were administered only to sample 2. Thus, the three categories of items were administered to three different groups of simulees with different ability distributions, which resulted in differences in the accuracy of item difficulty estimation. The RMSE($\hat{\beta}$) values of the items in the intermediate difficulty category were comparable to those found in the random condition. For the easy and difficult items, however, targeted assignment of simulees led to lower RMSE($\hat{\beta}$) values than random assignment. These results suggest that the targeted design outperformed the random condition by improving the calibration of easy and difficult items.

For TMST A, the item pool was divided into five difficulty categories, which resulted in a curve consisting of five sections. Categories 2 and 4 corresponded to the items in stage 1 (i.e., relatively easy and relatively difficult items). The RMSE($\hat{\beta}$) of these items were comparable to the outcomes of the random and the targeted conditions. However, TMST A provided more accurate item difficulty estimates for the items in stage 2. On the one hand, a reduction in RMSE($\hat{\beta}$) was achieved for easy and difficult items (i.e., categories 1 and 5). These items were administered to low performing simulees from sample 1 or high performing simulees from sample 2. On the other hand, RMSE($\hat{\beta}$) was also lower for intermediate items from category 3, which were administered to high performing simulees from sample 1 and to low performing simulees from sample 2. The efficiency gain in this intermediate category could be explained partly by the enhanced mean numbers of observations per item caused by the slightly unbalanced routing (M_3 = 447; $M_1 = M_5$ = 426). Given the composition of the modules in stage 1 and the limited score

range of 0–15, this outcome represented the best approximation of a balanced number of observations over the entire design.

Finally, TMST B included four difficulty categories. However, the curve was only divided into three sections such that the two middle categories related to stage 1 were not visually distinguishable. Similar to TMST A, TMST B provided more accurate item difficulty estimates for easy and difficult items (i.e., categories 1 and 4) than the targeted design or the random condition. These two difficulty categories corresponded to stage 2 of TMST B, where the simulees' assignment was determined by their performance in stage 1.

Taken together, the item-level findings provided further insights into the efficiency of the different designs under the full knowledge condition. The targeted design outperformed the random condition owing to greater accuracy in item difficulty estimation for the easy and the difficult items. The two targeted multistage designs further improved the item difficulty estimation for the easy and the difficult items compared to the targeted design by means of the performance-based assignment of simulees to items in stage 2.

Conditions with limited knowledge. Figure 8 shows six charts. Each chart presents RMSE($\hat{\beta}$) in relation to the true difficulty β for one of the three designs in combination with one of the two limited knowledge conditions. The random condition is included as the baseline in all six charts (black line). As an additional orientation, RMSE($\hat{\beta}$) under the full knowledge condition is displayed for each design as a gray line.

The two charts in the first row of Figure 8 refer to the targeted design, where we distinguished three difficulty categories. Within each difficulty category, the relationship between RMSE($\hat{\beta}$) and β was represented by a U-shaped curve. Differences in RMSE($\hat{\beta}$) were found mainly for easy and difficult items, but not for the items in the intermediate difficulty range.

Items that were assigned correctly to their corresponding difficulty category achieved similar RMSE($\hat{\beta}$) values as under the full knowledge condition. However, easy or difficult items that were placed wrongly in the intermediate category showed RMSE($\hat{\beta}$) values comparable to those under the random condition. Moreover, the placement of difficult items in the easy category and easy items in the difficult category resulted in RMSE($\hat{\beta}$) values that exceeded the values under the random condition. The number of misplaced items and, therefore, the number of items with high RMSE($\hat{\beta}$) values were lower under the medium knowledge condition (left chart) than under the low knowledge condition (right chart).

The two charts in the second row of Figure 8 refer to TMST A, where we distinguished five difficulty categories. Within each difficulty category, the relationship between RMSE($\hat{\beta}$) and β was again represented by a U-shaped curve. However, variations in RMSE($\hat{\beta}$) were clearly larger for TMST A under the limited knowledge conditions than for the targeted design. Considerable differences emerged in all five difficulty categories. In particular, the RMSE($\hat{\beta}$) values of the items in the intermediate difficulty category (i.e., category 3) were high. Even items of intermediate difficulty, which were assigned correctly to this category, showed RMSE($\hat{\beta}$) values greater than those under the random condition.

These findings were related to large variations in the mean number of observations in stage 2 of TMST A (see Table 1). Under the limited knowledge conditions, a substantial number of observations shifted away from the intermediate modules in stage 2 toward the easy and the difficult modules, which resulted in considerably lower mean numbers of observations for the items in the intermediate modules (medium knowledge: $M_3 = 312$; low knowledge: $M_3 = 252$). This shift was caused by the shift in the mean module difficulty in stage 1, as shown in Figure 6.

The relatively easy modules in stage 1 became more difficult due to misplaced items, which resulted in a lower percentage of simulees surpassing the cut-off score and reaching the intermediate modules in stage 2. Simultaneously, the relatively difficult modules became easier, such that a greater number of simulees surpassed the cut-off score and reached the difficult modules. Thus, the impaired overall efficiency of TMST A under the conditions of limited knowledge seems to be triggered mainly by the misplacement of a few items in stage 1 and the related impact on routing.

Finally, the two charts in the third row of Figure 8 refer to TMST B, where we distinguished four difficulty categories. In line with previous results, the relationship between RMSE($\hat{\beta}$) and β was represented by a U-shaped curve for each difficulty category. Similar to the targeted design, differences in RMSE($\hat{\beta}$) values were larger for easy and difficult items than for relatively easy or difficult items. Easy items assigned to the relatively easy category and difficult items assigned to the relatively difficult category were estimated more efficiently with TMST B than with the random condition. However, easy items assigned to one of the two difficult categories and difficult items assigned to one of the two easy categories resulted in RMSE($\hat{\beta}$) values higher than those observed under the random condition. For items assigned to the appropriate difficulty category, again, we found similar RMSE($\hat{\beta}$) values as under the full knowledge condition, with one exception: slightly higher RMSE($\hat{\beta}$) values were reported for items in the difficult category under the medium knowledge condition. This difference was caused by a slightly lower mean number of observations for the items in the difficult category. Nevertheless, we concluded that limited knowledge resulted in similar effects for TMST B and the targeted design, namely, lower accuracy for misplaced items.

Discussion

In this paper, we introduced the concept of targeted multistage calibration designs for calibrating CAT item pools dedicated to multiple ability groups and, thus, allowing for the measurement of ability over a broad range. We investigated the efficiency of this design type for calibrating items under the Rasch model by means of simulations. As expected, the two targeted multistage calibration designs were more efficient in calibrating the item pool than the targeted design or the random condition, given that complete knowledge about item difficulty was available during construction of the calibration designs. The reported improvements in overall RMSE corresponded to statistically significant, as well as practically relevant, efficiency gains in terms of the number of students. Namely, additional time and financial resources would be required to recruit up to 200 additional students for calibrating the items with the traditional targeted design in order to compensate for the gain of the targeted multistage calibration designs.

Differences in the accuracy of item difficulty between the targeted design and the two targeted multistage designs were found especially for easy and difficult items. Reliable item difficulty estimates for these groups of items are highly relevant in the context of CAT, where each student is assessed based on a different subset of items from the overall item pool (see also van der Linden & Glas, 2000). The performance-based routing in the targeted multistage designs allowed for identification of low and high performing students including students with extreme and, thus, rare abilities within both samples, and for assigning them specifically to the easy and difficult items. Therefore, the multistage procedure improved the match between item difficulty and ability for those items that clearly differed from the mean abilities of the two samples. In practice, such improved alignment between item difficulty and student ability is not only beneficial for calibrating items, but it also provides more reliable estimates of ability. Moreover, it

might prevent low and high ability students from getting discouraged or bored by items that do not fit their ability level (Asseburg & Frey, 2013). Sustaining student motivation during test taking is, in turn, an important basis for producing reliable calibration outcomes (e.g., Finn, 2015; Mittelhaëuser, Béguin, & Sijtsma, 2015; Zwitser & Maris, 2015).

However, the construction of efficient calibration designs requires knowledge not only about the abilities of calibration samples but also about item difficulty. Unfortunately, empirical knowledge about item difficulty is often not available in practical settings prior to calibration. Instead, difficulty ratings from experts are used as approximations of real item difficulty and as a basis for distributing items within a design. Several studies have reported medium-to-high correlations between such ratings and real item difficulties (e.g., Bejar, 1983; Hambleton & Jirka, 2006; Sydorenko, 2011; Wauters et al., 2012). Therefore, we investigated the extent to which misclassification of items in terms of their difficulty could impair the efficiency of the different calibration designs considered in this study. In line with our expectations, limited knowledge about item difficulty was related to a considerable decrease in the efficiency of the traditional targeted design as well as of the two targeted multistage calibration designs.

Furthermore, we hypothesized that targeted multistage calibration designs are more vulnerable to misplaced items than targeted calibration designs because knowledge about item difficulty is crucial for specifying appropriate routing rules. This hypothesis was only partly supported by the results of our simulation study. In line with our expectations, we found a severe loss of efficiency for one of the two targeted multistage calibration designs under the conditions of limited knowledge and misplaced items. The efficiency loss of TMST A was caused by a mismatch between the specified routing rules based on predicted item difficulties and the real mean difficulties of the routing modules. This mismatch led to inefficient distribution of simulees over

the modules of stage 2, considerably low mean numbers of observations for certain items, and, finally, relatively inaccurate item difficulty estimates of the affected items.

However, the second multistage calibration design was more robust for limited knowledge and misplaced items. Contrary to our expectations, TMST B outperformed the targeted design, regardless of the amount of available knowledge, even though the advantage of TMST B over the targeted design was greater under the full knowledge condition. The specific combination of routing module composition and routing rules in TMST B managed to compensate for changes in routing due to misplaced items. Nevertheless, the minor differences in the mean numbers of observations under the medium knowledge condition suggest that overlapping routing paths do not guarantee a stable targeted multistage calibration design. In our study, the mean difficulty of the routing modules shifted toward the mean of the item pool under the limited knowledge conditions. However, experts might also systematically under- or overestimate the difficulty of all items, leading to under- or overrepresentation of the numbers of observations for the easy and the difficult modules in stage 2. In contrast, targeted calibration designs provide full control over the assignment of students to test booklets and, thereby, over the number of observations per item—independent of the amount of knowledge about the difficulty of the items.

As in any simulation study, our study covered a limited set of conditions, which limits generalizability of the findings in some ways (Davey, Nering, & Thompson, 1997; Feinberg & Rubright, 2016). First, we investigated the efficiency of item calibration under the Rasch model. Different results might be found if the response patterns do not fit the Rasch model perfectly as they did in our simulation study. Our results could serve as a starting point for investigating the efficiency of targeted multistage calibration designs in combination with more complex IRT models. Based on previous studies, we hypothesize that targeted multistage designs allow for the

efficient estimation of item difficulty parameters independent of the IRT model (Berger, 1991; Stocking, 1988). However, further research is needed for investigating the extent to which the limited ability variation within each subsample could impair efficient estimation of other item parameters, such as discrimination or guessing.

Second, we focused on improving the match between item difficulty and student ability for enhancing calibration efficiency. However, Eggen and Verhelst (2006), as well as Verschoor (2010), found that the efficiency of incomplete calibration designs also depends on the strength of the links between the modules within a design so that designs with a larger number of linking items provide more accurate estimates of item difficulty. However, given the test length of 30 items and the use of MML estimation procedures, we expect only small improvements through such adaptations.

A third limitation of our study is that the size of the samples was relatively small. In alignment to a practical setting, we included 2,600 simulees in each simulation run. An enhancement of sample size would significantly increase the accuracy of item difficulty estimates under all conditions and decrease the practical relevance of efficiency gain through better alignment of item difficulty and student ability (Wright, 1977, p. 105). On the other hand, recruiting samples that are large enough for item calibration is often challenging, which underlines the practical relevance of this constraint. Moreover, it would be interesting to investigate in further studies whether the current results are replicable with more complex designs, different item pools, and different samples. Based on our findings, we hypothesize that targeted multistage designs will always outperform targeted designs under optimal conditions. However, we assume that the risk of efficiency loss in targeted multistage calibration designs depends not only on the amount of

available knowledge about item difficulty, but also on the particular combination of design specifications, item pool, and sample. Thus, further studies should analyze the factors that can contribute to the stability of targeted multistage calibration designs. Furthermore, it would be interesting to examine whether step-by-step adaption of the routing rules or of the composition of modules during calibration (cf. Ali & Chang, 2014; Kingsbury, 2009) could reduce the risk of efficiency loss due to imbalanced distribution of the number of observations per item.

Efficient calibration designs are crucial in practice because it is often difficult to recruit sufficient numbers of students for calibrating CAT item pools. With our simulation study, we extended previous research on the efficiency of calibration designs by introducing targeted multistage calibration design as a new design type and by investigating a practically relevant constraint, namely, the influence of limited knowledge during test development. We conclude from our findings that targeted multistage calibration designs are an option for enhancing the efficiency of calibration under the condition that reliable knowledge about item difficulty is available. However, these theoretically superior designs come at a certain price in practice if a priori knowledge about item difficulty is limited. Therefore, we recommend practitioners to use targeted designs instead of targeted multistage designs for calibration whenever there are doubts about the accuracy of predicted item difficulty.

References

Ali, U. S., & Chang, H.-H. (2014). *An item-driven adaptive design for calibrating pretest items* (ETS Research Reports Series No. RR-14-38). Princeton, NJ: Educational Testing Service.

Angelone, D., Keller, F., & Moser, U. (2013). *Entwicklung schulischer Leistungen während der obligatorischen Schulzeit: Bericht zur vierten Zürcher Lernstandserhebung zuhanden der Bildungsdirektion des Kantons Zürich* [Development of school performance during compulsory school: Report on the fourth assessment for the attention of the Zurich department of education]. Zürich: Institut für Bildungsevaluation. Retrieved from Institut für Bildungsevaluation website: http://www.bi.zh.ch/internet/bildungsdirektion/de/unsere_direktion/bildungsplanung/arbeitenundprojekte/lernstand/_jcr_content/contentPar/downloadlist_3/downloaditems/347_1408695301431.spooler.download.1408694967389.pdf/Lernstand9Klassen_Wissenschaftlicher+Bericht+2013.pdf

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, *55*, 92–104.

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, *7*, 303–310. doi:10.1177/014662168300700306

Berger, M. P. F. (1991). On the efficiency of IRT models when applied to different sampling designs. *Applied Psychological Measurement*, *15*, 293–306. doi:10.1177/014662169101500310

Berger, S., Moser, U., Verschoor, A. J., & Eggen, T. J. H. M. (2015). *Development of an online item bank for adaptive formative assessment*. Paper presented at the AEA-Europe Conference. AEA-Europe, Glasgow, Scotland.

Berger, S., Oostlander, J., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2015). *Development of multistage tests based on teacher ratings*. Paper presented at the conference of the International Association for Computerized Adaptive Testing, Cambridge, England.

Berger, S., Verschoor, A. J., Moser, U., & Oostlander, J. (2014). *Educational assessment in Northwestern Switzerland: Psychometrical challenges in linking standardized tests to an online item bank for formative assessment*. Paper presented at the AEA-Europe Conference, Tallinn, Estonia.

Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data* (ACT Research Report Series No. 97-4). Iowa City, IA: ACT. Retrieved from ACT website: http://www.act.org/content/dam/act/unsecured/documents/ACT_RR97-04.pdf

Eggen, T. J. H. M., & Verhelst, N. D. (2006). Loss of information in estimating item parameters in incomplete designs. *Psychometrika*, *71*, 303–322. doi:10.1007/s11336-004-1205-6

Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicológica*, *32*, 107–132.

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*, 36–49. doi:10.1111/emip.12111

Finn, B. (2015). *Measuring motivation in low-stakes assessments* (ETS Research Reports Series No. RR-15-19). Princeton, NJ: Educational Testing Service.

Glas, C. A. W. (2010). MIRT: Multidimensional item response theory. Enschede: University of Twente. Retrieved from https://www.utwente.nl/nl/bms/omd/Medewerkers/temp_test/mirt-manual.pdf

Glas, C. A. W., & Geerlings, H. (2009). Psychometric aspects of pupil monitoring systems. *Studies in Educational Evaluation*, *35*, 83–88. doi:10.1016/j.stueduc.2009.10.006

Hambleton, R. K., & Jirka, S. J. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 399–420). Mahwah, NJ: Lawrence Erlbaum Associates.

Kingsbury, G. G. (2009). Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing.* Retrieved from http://www.iacat.org/sites/default/files/biblio/cat09kingsbury.pdf

Kolen, M. J. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31–55). New York, NY: Springer.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York, NY: Routledge.

Makransky, G., & Glas, C. A. W. (2010). An automatic online calibration design in adaptive testing. *Journal of Applied Testing Technology*, *11*, 1–20.

Mislevy, R. J., & Wu, P. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (ETS Research Reports Series No. RR-96-30-ONR). Princeton, NJ: Educational Testing Service.

Mittelhaëuser, M.-A., Béguin, A. A., & Sijtsma, K. (2015). The effect of differential motivation on IRT linking. *Journal of Educational Measurement*, *52*, 339–358. doi:10.1111/jedm.12080

Northwest Evaluation Association. (2011). *Technical manual for Measure of Academic Progress & Measure of Academic Progress for Primary Grades*. Portland, Oregon. Retrieved from

https://www.richland2.org/RichlandDistrict/media/Richland-District/AdvancED/Standard%205/5.1/5-1-NWEA-Technical-Manual-for-MAP-and-MPG.pdf

O'Neill, T. R., Marks, C. M., & Reynolds, M. (2005). Re-evaluating the NCLEX-RN passing standard. *Journal of Nursing Measurement, 13*, 147–165. doi:10.1891/jnum.2005.13.2.147

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Renaissance Learning. (2015). *STAR Math: Technical manual*. Wisconsin Rapids, WI: Renaissance Learning.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion* [Textbook test theory, test construction]. Bern: Verlag Hans Huber.

Stocking, M. L. (1988). *Specifying optimum examinees for item parameter estimation in item response theory* (ETS Research Reports Series No. RR-88-57-ONR). Princeton, NJ: Educational Testing Service.

Strobl, C. (2012). *Das Rasch-Modell: Eine verständliche Einführung für Studium und Praxis* [The Rasch model: A coherent introduction for students and practitioners] (2nd ed.). Mering: Rainer Hampp Verlag.

Sydorenko, T. (2011). Item writer judgments of item difficulty versus actual item difficulty: A case study. *Language Assessment Quarterly, 8*, 34–52. doi:10.1080/15434303.2010.536924

Vale, C. D., & Gialluca, K. A. (1988). Evaluation of the efficiency of item calibration. *Applied Psychological Measurement, 12*, 53–67. doi:10.1177/014662168801200106

van der Linden, W. J. (1988). *Optimizing incomplete sample designs for item response model parameters* (Research Report No. 88-5). Enschede: University of Twente.

van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education, 13,* 35–53. doi:10.1207/s15324818ame1301_2

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York, NY: Springer.

Verschoor, A. J. (2010). *Optimal calibration designs for computerized adaptive testing*. Paper presented at the first annual conference of the International Association for Computerized Adaptive Testing, Arnhem, the Netherlands.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Wauters, K., Desmet, P., & van den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education, 58,* 1183–1193. doi:10.1016/j.compedu.2011.11.020

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

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21,* 361–375. doi:10.1111/j.1745-3984.1984.tb01040.x

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97–116. doi:10.1111/j.1745-3984.1977.tb00031.x

Yan, D., von Davier, A. A., & Lewis, C. (Eds.). (2014). *Computerized multistage testing: Theory and applications*. Boca Raton, FL: CRC Press.

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). New York, NY: Springer.

Zwitser, R. J., & Maris, G. (2015). Conditional statistical inference with multistage testing designs. *Psychometrika*, *80*, 65–84. doi:10.1007/s11336-013-9369-6

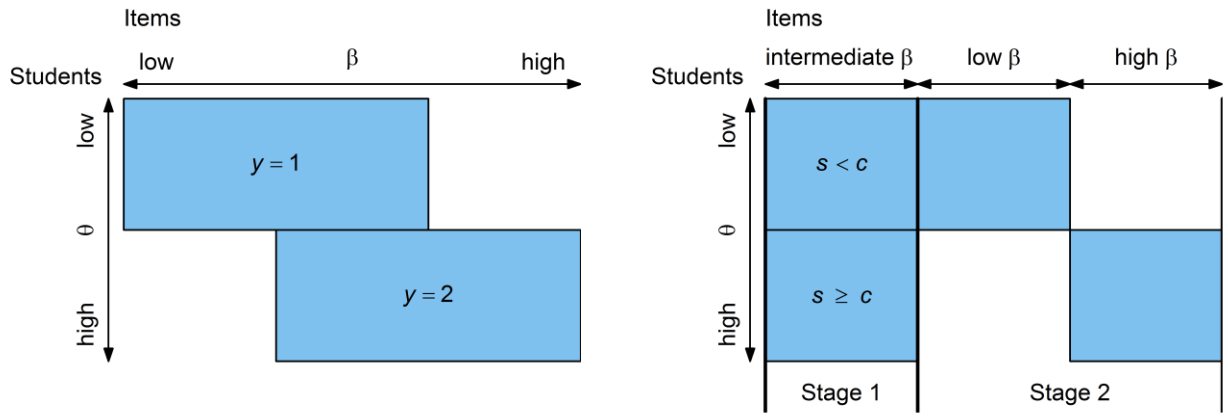

Figure 1. Examples of targeted (left) and multistage calibration designs (right). y = ability-related background variable (e.g., grades in school); s = student's score in stage 1, c = cut-off score.

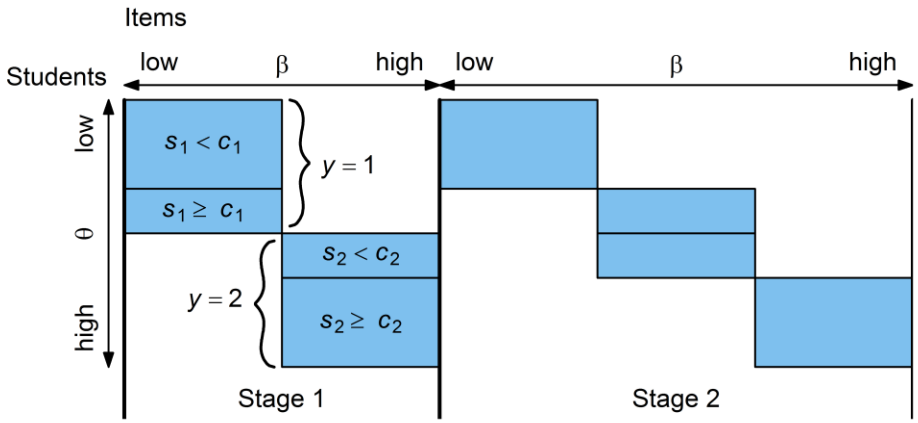

Figure 2. Example of targeted multistage calibration design. y = ability-related background variable (e.g., grades in school); s_1 and s_2 = student's scores in stage 1, c_1 and c_2 = cut-off scores.

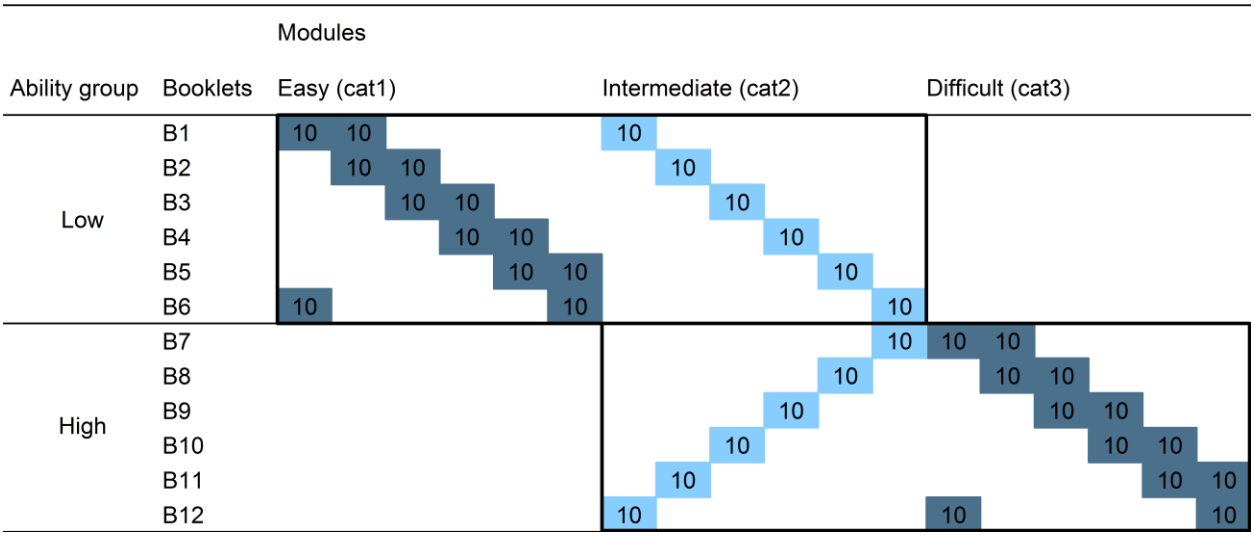

Figure 3. Targeted calibration design. B1–B6 = easy booklets; B7–B12 = difficult booklets. Each column represents one module of ten items. The 18 modules are classified into three difficulty categories: cat1 = “easy,” cat2 = “intermediate,” and cat3 = “difficult.”

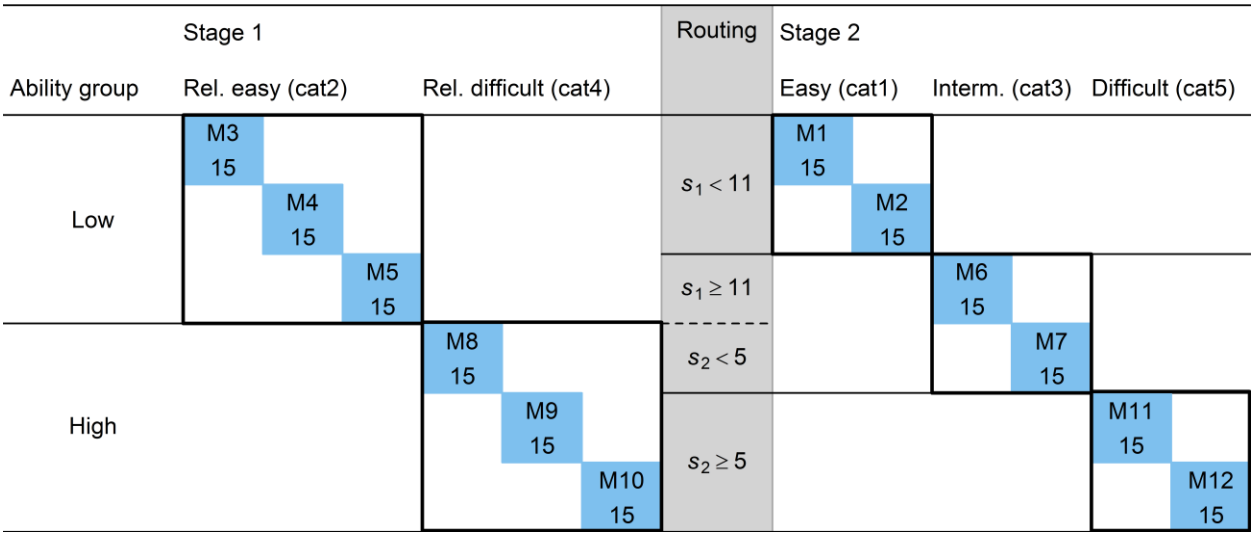

Figure 4. Targeted multistage calibration design A (TMST A). M1/M2 = easy modules (difficulty category 1); M3–M5 = relatively easy modules (category 2); M6/M7 = intermediate modules (category 3); M8–M10 = relatively difficult modules (category 4); M11/M12 = difficult modules (category 5); s_1 = score of low ability group; s_2 = score of high ability group.

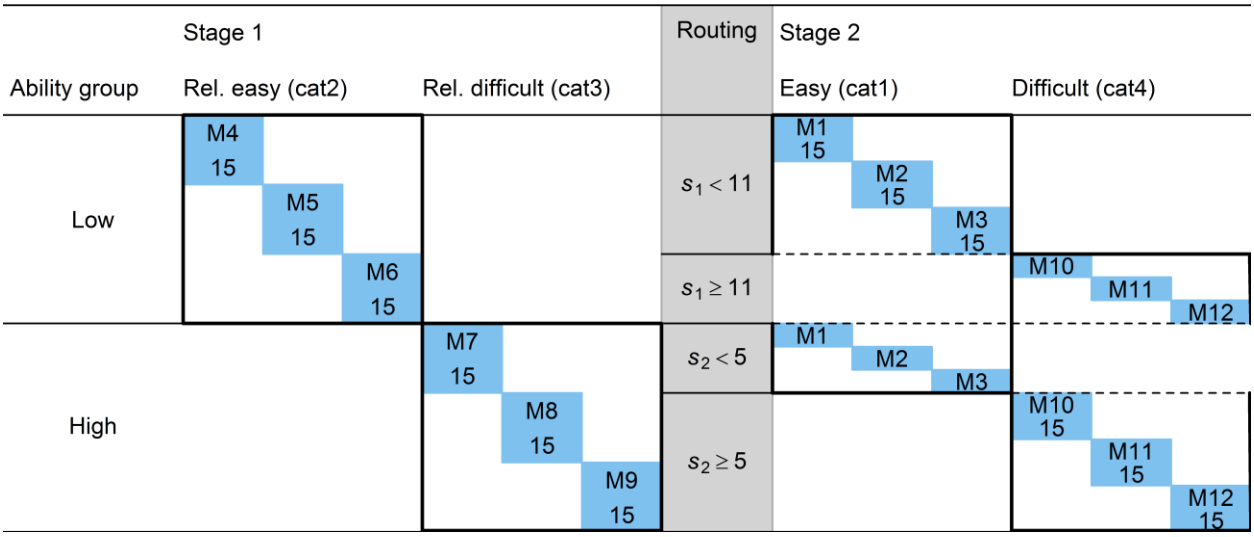

Figure 5. Targeted multistage calibration design B (TMST B). M1–M3 = easy modules (difficulty category 1); M4–M6 = relatively easy modules (category 2); M7–M9 = relatively difficult modules (category 3); M10–M12 = difficult modules (category 4); s_1 = score of low ability group; s_2 = score of high ability group.

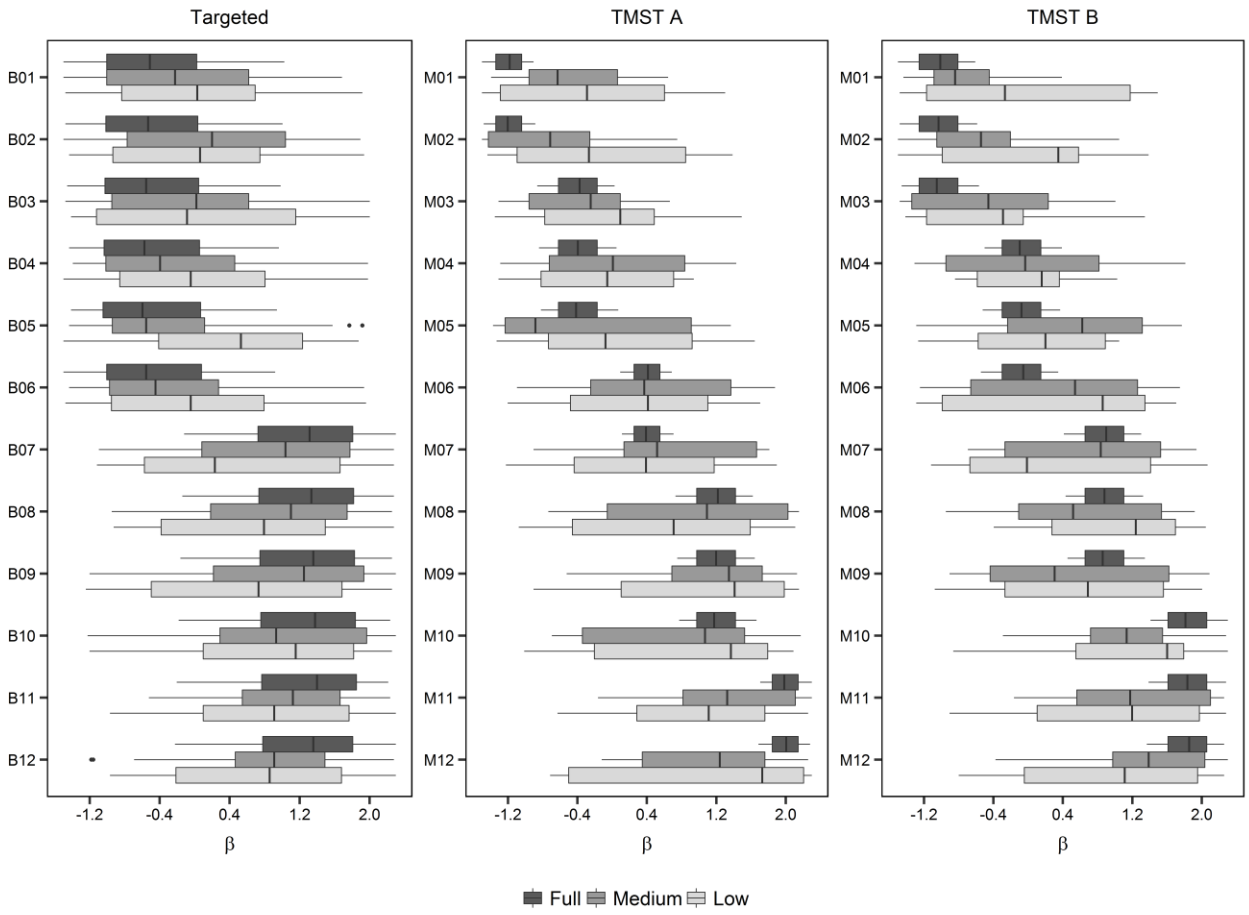

Figure 6. Boxplots of item difficulty per design, booklet/module, and knowledge condition.

Table 1.

Range of Bias($\hat{\beta}$), Mean and Standard Error of RMSE($\hat{\beta}$), Related Gain per Simulation Condition, and Distribution of Mean Number of Observations

	Bias($\hat{\beta}$)[a]		RMSE($\hat{\beta}$)		Gain over Random		Gain over Targeted[c]		*N*					
Condition	*Min*	*Max*	*M*	*SE*[b]	%	*N* (%)	%	*N* (%)	*SD*	M_1	M_2	M_3	M_4	M_5
Random	–0.011	0.012	0.119	0.00028	--	--	--	--	0.549	433	--	--	--	--
Full Knowledge														
Targeted	–0.007	0.007	0.115	0.00027	3	159 (6)	--	--	0.222	433	433	433	--	--
TMST A	–0.005	0.005	0.112	0.00026	6	267 (10)	2	122 (5)	6.958	426	433	447	433	426
TMST B	–0.006	0.007	0.110	0.00026	7	336 (13)	4	200 (8)	0.114	433	433	433	433	--
Med. Knowledge														
Targeted	–0.007	0.007	0.117	0.00028	1	64 (2)	--	--	0.175	433	433	433	--	--
TMST A	–0.007	0.007	0.120	0.00029	**–1**	**–67 (–3)**	**–3**	**–138 (–5)**	60.958	492	433	312	433	496
TMST B	–0.007	0.006	0.116	0.00027	2	114 (4)	1	52 (2)	6.315	442	433	433	424	--
Low Knowledge														
Targeted	–0.006	0.007	0.119	0.00028	0	4 (0)	--	--	0.145	433	433	433	--	--
TMST A	–0.010	0.010	0.125	0.00030	**–6**	**–315 (–12)**	**–6**	**–321 (–12)**	90.950	519	433	252	433	529
TMST B	–0.008	0.007	0.118	0.00028	1	48 (2)	1	43 (2)	2.925	429	433	433	437	--

Note. The numbers in bold represent efficiency loss compared to the baseline condition; gray numbers represent mean numbers of observations equal to the overall mean of $M = 433$. N(%) = number of additional students needed to achieve the same efficiency with the baseline design, 100% = 2,600 students; M_1–M_5 = mean numbers of observations per difficulty category (i.e., design-specific categorization of modules' difficulty levels).

[a]Mean of Bias($\hat{\beta}$) was zero in all conditions.

[b]Standard error of RMSE($\hat{\beta}$) based on the first 1,000 simulation runs.

[c]Gain over targeted design within a specific knowledge condition.

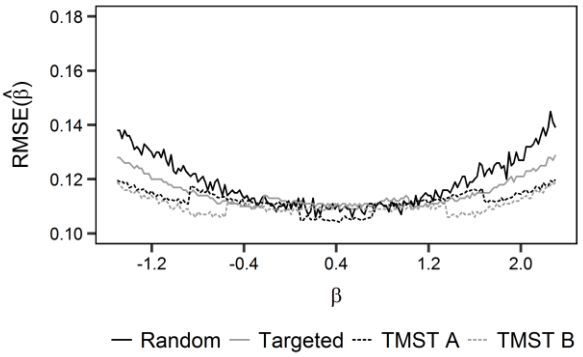

Figure 7. RMSE($\hat{\beta}$) per item in relation to item difficulty for the four different designs under the full knowledge condition.

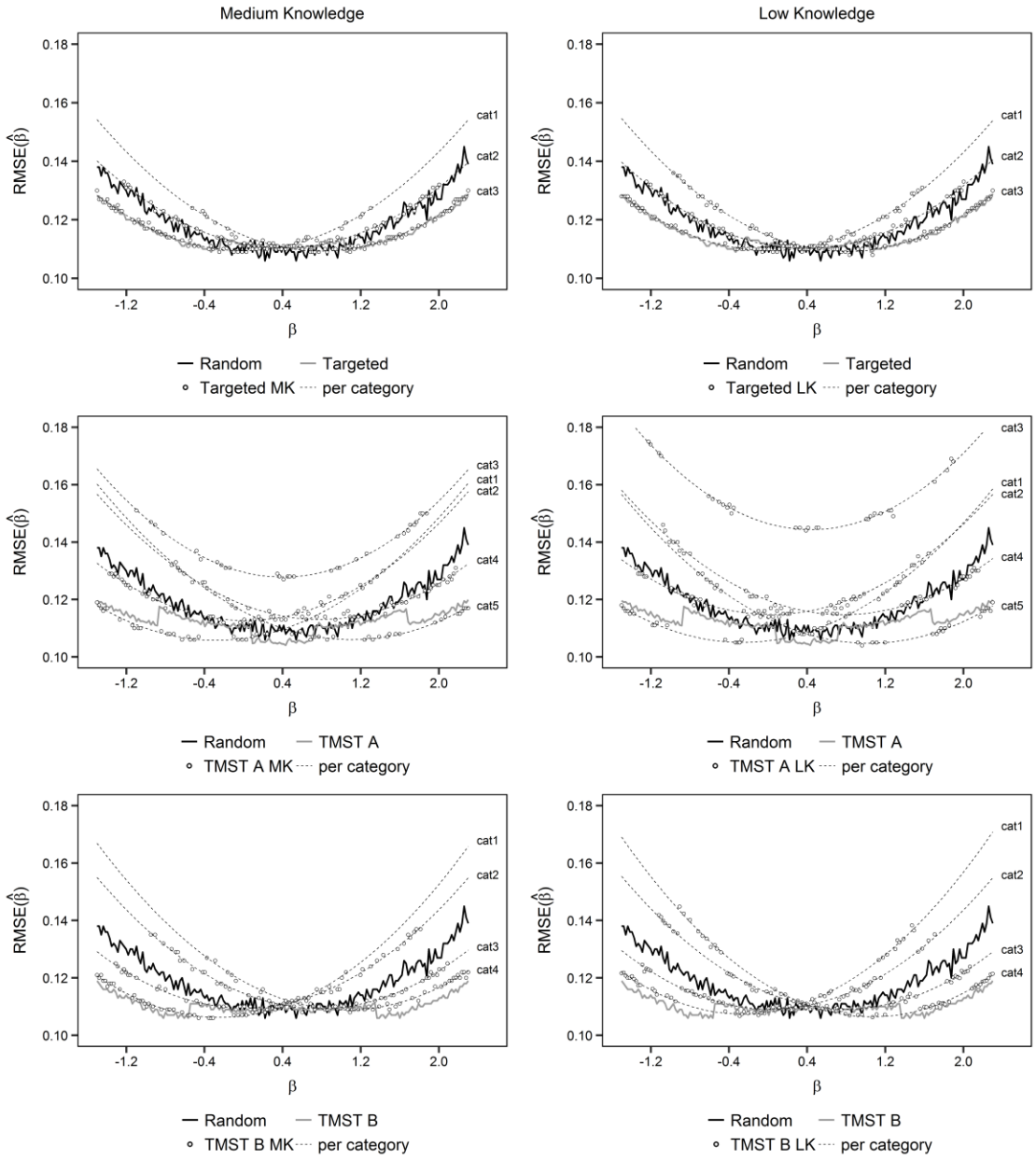

Figure 8. RMSE($\hat{\beta}$) per item in relation to item difficulty for the three different designs under the medium (left) and low knowledge (right) conditions. The solid black line denotes the random condition (baseline) and the solid gray line the optimal condition specific to each chart. The white dots refer to the limited knowledge conditions (LK = low knowledge, MK = medium knowledge), and the related dotted lines cat1 to cat5 indicate the regressed trends of each difficulty category under the limited knowledge conditions.