Psychological Test and Assessment Modeling, Volume 60, 2018 (3), 327-346

A continuous calibration strategy for computerized adaptive testing

Aron Fink[1], Sebastian Born[2], Christian Spoden[2,3] & Andreas Frey[2,4]

Abstract

This paper presents a new continuous calibration strategy for using computerized adaptive testing in application areas where it is not feasible to conduct a separate calibration study and/or to construct the complete item pool before the operational phase of the test. This method enables a step-by-step build-up of the item pool across several test cycles. A combination of equating and linking is used to maintaining the scale across these cycles. A simulation study was carried out to investigate the performance of the strategy regarding the precision of the ability estimates. The simulation study is based on a full factorial design with the factors IRT model, sample size and number of new uncalibrated items added to the item pool per test cycle. Precision of the ability estimates increased over the test cycles in all conditions. For the 2PL model, a better performance was reached when using a lower number of new uncalibrated items. The results support the application of the new method especially in small sample sizes.

Keywords: computerized adaptive testing, item response theory, online calibration, item banks, test design

[1] *Correspondence concerning this article should be addressed to:* Aron Fink, Institute of Educational Science, Department of Research Methods in Education, Friedrich Schiller University Jena, Am Planetarium 4, 07743 Jena, Germany, email: aron.fink@uni-jena.de

[2] Friedrich Schiller University Jena, Germany

[3] Now at the German Institute for Adult Education – Leibniz Centre for Lifelong Learning, Bonn, Germany

[4] Centre for Educational Measurement (CEMO) at the University of Oslo, Norway

Computerized adaptive testing (CAT) is a testing mode in which the selection of the item to be presented next to the test taker depends upon the responses given to previously administered items (Frey, 2012). Extensive prior research shows that CAT typically yields more precise ability estimates and/or a shorter test length compared to traditional nonadaptive testing (e.g., Segall, 2005). In addition, CAT has a great potential to overcome some limitations of traditional nonadaptive tests, which could explain its growing popularity in many different fields, such as psychological and educational testing, large-scale assessments, admission testing, health outcome assessments and others (for a list of operational CAT programs see http://www.iacat.org/content/operational-cat-programs). In particular, CAT can be designed so that test takers are measured with a comparable level of precision across the complete ability range (Frey & Ehmke, 2007). In contrast, traditional nonadaptive tests typically provide the highest precision for test takers of medium ability, while the precision decreases for test takers with test scores in the extremes (Dolan & Burling, 2012).

An essential building block in the development of a computerized adaptive test is a calibrated item pool (e.g., He & Reckase, 2014; Thompson & Weiss, 2011). Traditionally, one single calibration study is carried out, in which a large number of test takers respond to a large number of candidate items. Based on the responses gathered in the calibration study, item parameters are estimated by means of item response theory (IRT; e.g., van der Linden, 2016) methods. In the subsequent operational phase of a computerized adaptive test, the estimated item parameters are considered to be known and are used for item selection and ability estimation. Therefore, the quality of a computerized adaptive test depends, inter alia, on the precision of the item parameter estimates, which will, in turn, be determined mainly by the number of responses per item (i.e., the calibration sample size and the test length). For the frequently used two-parameter logistic (2PL) test model, for example, a minimum of 500 responses per item is recommended (de Ayala, 2009). Unfortunately, in various potential application areas of CAT, constructing large numbers of items prior to the initial use of the test (see Spoden, Frey and Bernhardt, in press, for a special case where using already field tested items was an option) and/or carrying out a calibration study with a large sample is not feasible, due to a lack of resources available for test development. Correspondingly, for applications such as written exams, psychological tests used in personnel selection, for clinical diagnosis or in research, CAT is typically not used, even though it would be advantageous here as well. In addition, test developers often access test takers from low-stakes samples for the calibration study, even if the intended population of test takers comes from another context. This procedure comes with a limitation. There is evidence that motivational differences exist between the low-stakes calibration samples and the intended high-stakes population of test takers (e.g., Sundre & Kitsantas, 2004; Wise & DeMars, 2005). These motivational differences may result in biased item parameter estimations in the calibration phase, which can jeopardize the interpretation of the test scores.

A possible solution to overcome these problems is the method of online calibration (e.g., Stocking, 1988; Kingsbury, 2009; Chen, 2017). Originally, online calibration methods were developed to pretest and calibrate new items on the fly during the administration of a computerized adaptive test. Most of the existing online calibration methods are, however, based on the assumption that an operational computerized adaptive test already exists and online calibration is only used to add new items to the item pool. Until now, the paper of Makransky and Glas (2010) is the only available study that uses online calibration designs to calibrate an item pool without having any information about the item parameters at the beginning of the operational CAT phase. Makransky and Glas (2010) presented a continuous updating calibration strategy, in which item parameters are re-estimated after each item administration using all previous responses. Therefore, the precision of the item and thus the precision of the ability estimates are continuously improved.

This strategy can be problematic for two major reasons. First, especially for high-stakes testing situations, re-estimating item parameters after each item administration leads to legal problems because ability measures are estimated from different item parameters and cannot be compared on the same established scale. This strategy is sound from a psychometric point of view but would be hard to defend in the court in case of a lawsuit. Second, in some cases the scales have to be linked across test cycles in order to make it possible to directly compare the test results across test cycles.

Against this background, a calibration strategy is necessary that allows for a simultaneous item and ability estimation and ensures a fair comparison of measures across test takers from different test cycles. The aim of the present study is thus to propose and examine a new strategy to calibrate items continuously across several test cycles without a separate calibration study, increasing the item pool across test cycles and maintaining the scale in each of these cycles.

Model

The present study was carried out in the framework of unidimensional logistic response models for dichotomous items. This family of models has a long tradition and is widely used in the field of educational and psychological testing (van der Linden, 2016). One model used in this study is the two-parameter logistic (2PL) model. This model defines the probability of a correct response $u_{ij} = 1$ of person $j = 1, \dots, N$ with the latent ability level θ_j to an item i as:

$$P\left(u_{ij} = 1 \middle| \theta_j, a_i, b_i\right) = \frac{\exp\left[\, a_i\left(\theta_j - b_i\right)\right]}{1 + \exp\left[\, a_i\left(\theta_j - b_i\right)\right]}, \tag{1}$$

where a_i is the discrimination parameter and b_i is the difficulty parameter of item i. Equation (1) reduces to the one-parameter logistic (1PL) model when the discrimination parameter a_i is set to a constant for all items (van der Linden, 2016). It should be noted that especially in high-stakes testing situations with closed response formats, the assumption of nonguessing, and thus a pseudoguessing parameter equal to zero is rather strict. However, the 2PL and the 1PL model are viable alternatives to more complex models because estimation of additional item parameters (e.g., pseudoguessing parameter) can be troublesome in small samples and tends to be less stable across different assessments. Unstable item parameter estimates complicate the process of linking across different test cycles, which makes it difficult to compare ability measures across these test cycles. Thus, the 2PL and the 1PL model are used in this study.

Continuous calibration strategy

The proposed continuous calibration strategy is divided into two phases, the *initial phase* and the *continuous phase*. The initial phase describes the first test cycle. During the initial phase, the same set of items is administered to every test taker. In fact, there is no difference between the initial phase of this procedure and a traditional nonadaptive test. Based on the item responses, item parameters and person parameters are estimated.

All test cycles following the initial phase are subsumed under the continuous phase. The major difference between the two phases is the type of item administration, which is nonadaptive in the initial phase and partly adaptive in the continuous phase. Tests in the continuous phase consist of three item clusters named *adaptive cluster*, *calibration cluster*, and *linking cluster*. The adaptive cluster contains items that are administered adaptively using the item parameter estimates from the previous test cycle. In the calibration cluster, items with unknown item parameters are included to enlarge the item pool. These new items are administered to every test taker of the current test cycle. Finally, the linking cluster comprises items that have already been administered in previous test cycles and have the most accurate item parameter estimates, compared to the other items in the item pool. These items are used to link consecutive test cycles with each other, using a common-item nonequivalent group design (Kolen & Brennan, 2014). The linking cluster makes it possible to report the results obtained in the various test cycles on the same scale and thereby allows direct comparisons of test results across these cycles. As a rule of thumb, at least 20 % of the test items should be used for linking, and each cluster of link items should be a good representation of the total test, both in content and statistical characteristics (Kolen & Brennan, 2014). To establish a stable linking between consecutive test cycles, it is necessary that the parameters of the link items are invariant across these cycles. Therefore, checking the link items for item parameter drift is necessary (IPD; Goldstein, 1983; Bock, Muraki, & Pfeiffenberger, 1988; Wells, Subkoviak, & Serlin, 2002). IPD occurs when the invariance assumption of the item parameters no longer holds across two or more test administrations. Link items showing significant differences in their item parameter estimates between two test cycles have to be identified and excluded from the linking procedure.

Each test cycle in the continuous phase can be divided into seven steps that are illustrated in Figure 1. First, items for the linking cluster of the current test cycle are selected (step 1). Subsequently, a test that is composed of an adaptive, a calibration and a linking cluster is administered (step 2). To check link items for IPD, item parameters are initially estimated based on the responses of the current test cycle (step 3). Since the examinees of the previous and current test cycles are not considered to be equivalent, parameter estimates for the two estimations are not on the same scale and therefore not directly comparable. Thus, a scale transformation has to be conducted. The link item parameters obtained from step 3 are brought onto the same scale as their parameters from the preceding test cycle by means of scale transformation methods (step 4; e.g., Kolen & Brennan, 2014). Thereafter, the set of link item parameter estimates is checked for IPD (step 5). Link items showing significant drift are excluded from linking and estimated freely in step 6a/6b. The test for IPD is accomplished iteratively (iterative purification method; Wells, Hambleton, Kirkpatrick, & Meng, 2014). After excluding link items showing IPD, the scale transformation of the remaining link items (step 4) has to be repeated again. This could lead to different results of the test for IPD, and, if necessary, to the exclusion of additional link items. The iterative process stops whenever there is no longer a link item showing significant IPD. Subsequently, item parameters are estimated using a fixed item parameter calibration based on the responses gained from all previous test cycles, whereby the item parameters of the remaining link items from step 5 are fixed at their estimates from the preceding test cycle (step 6b). Whenever a complete breakdown of the link occurs (i.e., there are not enough link items left to establish a stable linking), all item parameters are estimated freely based on the responses gained from all previous test cycles (step 6a). This type of estimation is often referred to as concurrent calibration (Wingersky & Lord, 1984). In this case, the results are not on the same scale as the results of the previous test cycles anymore, and, therefore, the linking procedure has to be started anew. Note that at least two link items need to be stable to keep the location and variation of the scale comparable across test cycles. However, it should be noted that the fewer the items used for the linking, the more prone to sampling errors the linking procedure is (Wingersky & Lord, 1984). Therefore, test developers should wisely choose the minimum number of items needed for fixed item parameter calibration. Finally, the ability parameters are estimated based on the item parameters from step 6a/6b (step 7).

To sum up, by applying the continuous calibration strategy, the item pool size increases by adding new items to the item pool, the item parameters are updated continuously while maintaining the reporting scale by means of the linking procedure, and the test gains precision and adaptivity across test cycles. In addition, differentiating between the three item clusters adds a lot of flexibility to the algorithm. If increasing the item pool size is not a primary interest anymore, the calibration cluster becomes an additional adaptive cluster and, therefore, adaptivity is maximized.

The continuous calibration strategy in its basic form combines different psychometric approaches and therefore has many tuning parameters (e.g., sample size per test cycle, underlying IRT model, test length, number of items per cluster, link item selection method, test for IPD). To find the best configuration of the proposed strategy and to give practical recommendations, these tuning parameters should be examined prior its operational use.

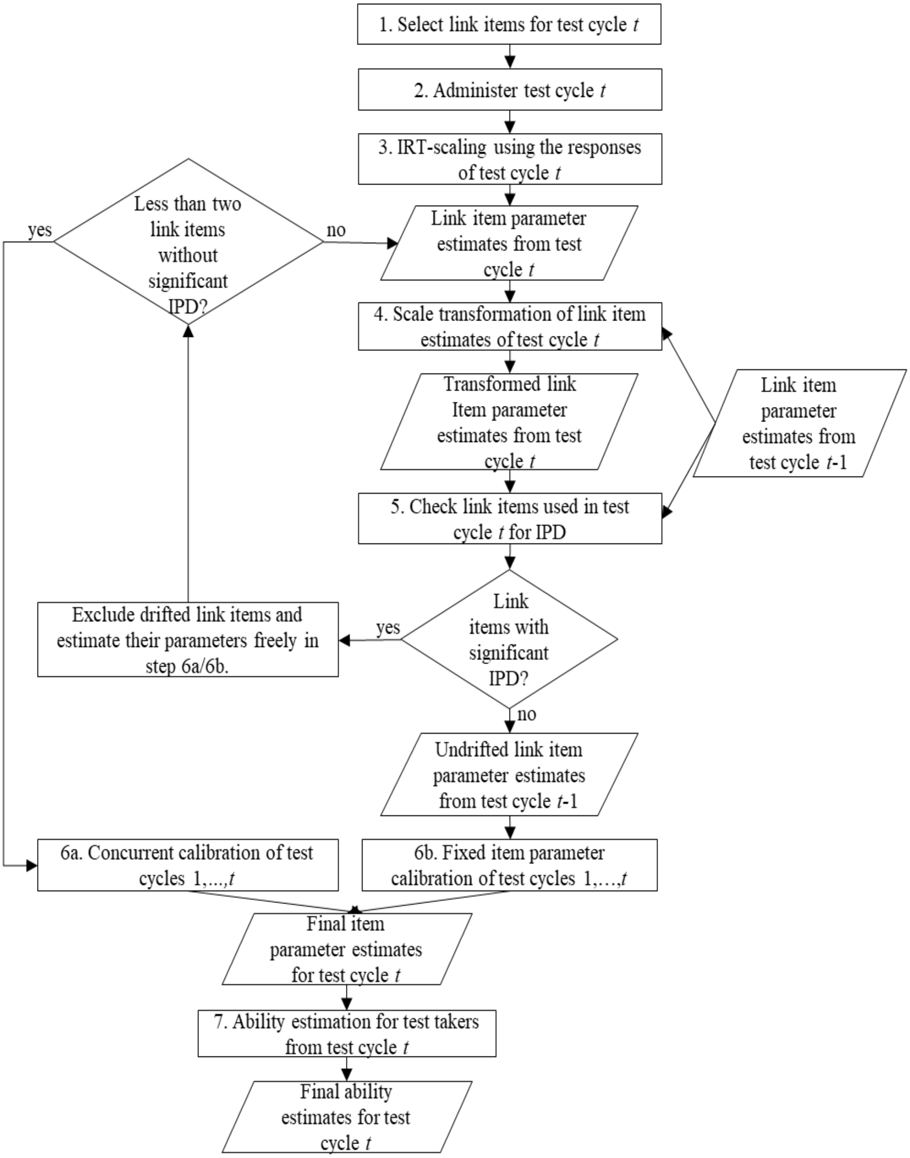
1. Select link items for test cycle *t*
2. Administer test cycle *t*
3. IRT-scaling using the responses of test cycle *t*
Link item parameter estimates from test cycle *t*
4. Scale transformation of link item estimates of test cycle *t*
Transformed link Item parameter estimates from test cycle *t*
Link item parameter estimates from test cycle *t*-1
5. Check link items used in test cycle *t* for IPD
Link items with significant IPD?
yes
no
Exclude drifted link items and estimate their parameters freely in step 6a/6b.
Less than two link items without significant IPD?
yes
no
Undrifted link item parameter estimates from test cycle *t*-1
6a. Concurrent calibration of test cycles 1,...,*t*
6b. Fixed item parameter calibration of test cycles 1,…,*t*
Final item parameter estimates for test cycle *t*
7. Ability estimation for test takers from test cycle *t*
Final ability estimates for test cycle *t*

Figure 1:
Flowchart of the continuous phase of the proposed continuous calibration strategy.

Research questions

The main aim of the following simulation study is to investigate the performance of the strategy regarding the quality of the obtained ability estimates in general and to give practical recommendations for the configuration of the algorithm. As there are numerous different possible configurations of the continuous calibration strategy, in this study only some of the basic tuning parameters are varied while others are kept constant. The first one is the sample size per test cycle. As already mentioned above, the number of test takers has a strong influence on the precision of the item parameter estimates. However, there are different recommendations concerning the required sample size for item calibration (e.g., “a few hundred”; de Ayala, 2009, p. 43). As the continuous calibration strategy is supposed to be applicable in areas where only small sample sizes are available for calibration purposes, the effect of different, rather small sample sizes on the performance of the continuous calibration is of interest. Therefore, the first research question is as follows:

1. What is the effect of using different sample sizes per test cycle in the continuous calibration strategy on the precision of the ability estimates at a different number of test cycles?

The second tuning parameter to be varied in this study is the number of items in the calibration cluster (which in turn means the number of test cycles necessary to obtain an initial estimation for each item in an item pool of a prespecified size), labeled *calibration speed.* This tuning parameter is supposed to strongly influence the performance of the continuous calibration strategy. Given a predefined test length, a higher number of items in the calibration cluster limits the number of items in the adaptive cluster. In addition, this leads to a faster increase in the number of items the selection algorithm can choose from. Thus, the level of calibration speed determines the level of adaptivity in each test cycle. In contrast, enlarging the item pool slowly leads to a more uniform item administration, which leads to a higher number of responses per item, and this in turn leads to more precise parameter estimates. The effect of using different levels of calibration speed is thus not easily predictable. Consequently, the second research question is:

2. What is the effect of using different levels of calibration speed in the continuous calibration strategy on the precision of the ability estimates at a different number of test cycles in the calibration process?

The last tuning parameter to be varied in this study is the underlying IRT model. As already mentioned above, in this study the 1PL and the 2PL model are used. Since the discrimination parameters in the 2PL model tend to be less stable than the difficulty parameters, a larger number of responses per item is recommended for calibration compared to the 1PL model. Bearing in mind that the number of responses per item in this study depends on the sample size and the calibration speed, it can be expected that there are differences in the results regarding research questions 1 and 2 when using different IRT models. Therefore, the last research question is:

3. To what extent does the chosen IRT model in the continuous calibration strategy affect the results for research questions 1 and 2?

Method

A study design with three independent variables was used. With the first independent variable, *Sample Size*, the number of test takers per test cycle ($N = 50$, $N = 100$, $N = 300$) was varied. The second independent variable, *Calibration Speed* ($t = 3$, $t = 5$, $t = 9$), compares the number of test cycles t necessary to obtain an initial estimation for each item in the item pool. Please note that concerning the calibration speed, the size of the item pool to be calibrated was 130 items, the test length for one test cycle was set to 50 items, and the number of items in the linking cluster for one test cycle was set to 10 items. For $t = 9$, every test in the continuous phase comprised 30 items in the adaptive cluster and 10 items in the calibration cluster. Since the item pool increased by 10 items in each test cycle, the complete item pool was calibrated after nine test cycles. This condition represents the slowest calibration procedure used in this study and determines the number of test cycles the continuous calibration was running in each condition. Thus, in order to keep the overall number of responses constant across conditions, the number of test cycles was set to nine for each condition. For $t = 5$, the tests in the continuous phase contained 20 items in the adaptive cluster and 20 items in the calibration cluster. Thus, after five test cycles every item had an initial estimation. From test cycle six to nine the calibration cluster became an additional adaptive cluster. For $t = 3$, the tests in the continuous phase comprised 10 items in the linking cluster and 40 items in the calibration cluster. No items were administered adaptively until test cycle four. From test cycle four to nine, the calibration cluster became the adaptive cluster. This condition represents the fastest possible calibration procedure, given the test specifications. The third independent variable, *IRT Model* (1PL, 2PL), represents the underlying IRT model used for calibration.

The fully crossed design had 2 x 3 x 3 = 18 conditions. For each of the conditions, 200 replications were analyzed with regard to the measurement precision as subsequently defined.

The simulation was carried out in R (R Core Team, 2017) using the “mirtCAT” package (Chalmers, 2016) for simulating the adaptive tests and the “mirt” package (Chalmers, 2012) for item and person parameter estimation. These functions were called from R-code that was written anew to carry out the continuous calibration strategy.

Simulation procedure

For each replication, the ability parameters were randomly drawn from a standard normal distribution, $\theta \sim N(0,1)$. The b_i parameters for each replication were drawn from a truncated normal distribution $b_i \sim N(0,1.5)$, $b_i \in (-4.5,4.5)$. This kind of distribution was chosen, because it can be assumed that under real conditions item parameters are usually located within this interval. The a_i parameters were drawn from a lognormal distribution, $a_i \sim lognormal(0,0.25)$.

Items in the calibration cluster and the linking cluster were administered sequentially; items in the adaptive cluster were selected using the maximum information criterion (Lord, 1980) based on the item parameter estimates obtained in the preceding test cycle.

After each test cycle, the items were calibrated under either the 1PL or the 2PL model using marginal maximum likelihood (MML, Bock & Aitkin, 1981) estimation and subsequently, person parameters were estimated using weighted maximum likelihood estimation (WLE; Warm, 1989). This method was preferred over maximum likelihood estimation, because it is less biased and provides ability estimates for test takers with invariant response patterns.

Selection of link items

For the selection of link items, previously administered items were categorized based on their estimated difficulty parameters. The interval limits of the categories were determined as quantiles of the item difficulty distribution:

Category 1 (very low difficulty): $b_i \in (b_{min}, b_{.1}]$;
Category 2 (low difficulty): $b_i \in (b_{.1}, b_{.3}]$;
Category 3 (medium difficulty): $b_i \in (b_{.3}, b_{.7}]$;
Category 4 (high difficulty): $b_i \in (b_{.7}, b_{.9}]$;
Category 5 (very high difficulty): $b_i \in (b_{.9}, b_{max})$.

Within each of these five categories, items with the lowest standard error (*SE*) of the difficulty parameter estimate were selected as link items. One item from category 1, two items from category 2, four items from category 3, two items from category 4 and one item from category 5 were selected to serve as potential link items. This procedure ensured that the distribution of the link items resembled the distribution of the complete item pool and, therefore, the linking cluster was a good representation of the whole test in terms of statistical characteristics. Before using these items for the linking, they were tested for IPD (s. Figure 1). For this purpose, item parameter estimates obtained from step 3 of the continuous calibration strategy were brought up to the same scale as their parameter estimates from the preceding test cycle. There are several popular transformation methods that can be used: mean/mean (Loyd & Hoover, 1980), mean/sigma (Marco, 1977), item characteristic curves (Haebara, 1980), and test characteristic curves method (Stocking & Lord, 1983). In this study, the mean/mean method was chosen due to the simple and user-friendly implementation of the method in comparison to the characteristic curve methods. In addition, the mean/mean method was preferred over the mean/sigma method, because means are typically more stable than standard deviations, as noted by Baker and Al-Karni (1991), and mean/sigma ignores the information from the a-parameters for the 2PL model. It should be noted that if the test is suspected to have very atypical item parameter estimates, the use of the moment methods (mean/mean, mean/sigma) may be questionable since they are highly sensitive to item parameter outliers. For practitioners, Hanson and Béguin (2002) suggested that it would be beneficial to apply multiple linking procedures and compare the scaling results. However, for the purpose of this study, it is sufficient to apply only one easily implementable transformation method. After scale transformation, it was examined whether the item parameter estimates of a link item from the preceding test cycle falls into the 95 % confidence interval around the same item's parameter estimates from the current test cycle. If the difficulty parameter estimate, the discrimination parameter estimate or both fell outside the confidence interval, the assumption of item parameter invariance was rejected for this item and, consequently, this item was not used for fixed item parameter calibration. In addition to this relatively simple and easily implementable method, there are also more complex methods for detecting IPD (e.g., mixed distribution IRT models; Park, Lee, & Xing, 2016), which, however, set stronger demands on the data, especially regarding the required sample size. The test for IPD was done iteratively. Drifted link items were excluded from linking and estimated freely without imposing any constraints.

Evaluation criteria

The mean squared error (MSE) of the ability estimates was used to evaluate the precision of the ability estimation after each test cycle for each of the conditions. The global MSE was computed as the average squared difference between the ability estimates $\hat{\theta}_j$ and the true ability θ_j for the *N* test takers in each test cycle and was averaged across the *rep* = 200 replications:

$$MSE = \frac{1}{rep * N} \sum_{r=1}^{rep} \sum_{j=1}^{N} (\hat{\theta}_j - \theta_j)^2. \tag{2}$$

Low MSE indicates high measurement precision. In addition to the global MSE (Equation 2), the precision of the ability estimates given a particular ability level was also investigated. This conditional MSE was calculated for a specific range of ability measures after each test cycle. The seven ability ranges used for this procedure were $\theta_j \in (-Inf, -2]$, $\theta_j \in (-2, -1]$, $\theta_j \in (-1, 0.25]$, $\theta_j \in (-0.25, 0.25]$, $\theta_j \in (0.25, 1]$, $\theta_j \in (1, 2]$, and $\theta_j \in (2, Inf)$.

Furthermore, a lower (L) and an upper baseline (U) were simulated to compare the continuous calibration design to additional criteria. For the lower *MSE* baseline, $\hat{\theta}_j$ was estimated after each test cycle using the real item parameters to simulate the hypothetically best possible estimation without errors stemming from the estimation of the item parameters. The precision of $\hat{\theta}_j$ stemming from item parameters that were fixed after their first estimation and thus not updated continuously was set as the upper *MSE* baseline.

Results

Global precision

To answer the research questions, in a first step the global MSE after each test cycle was analyzed. Figure 2 shows the MSE across test cycles under the 1PL for all sample sizes and calibration speeds. As seen, the MSE decreased over the test cycles in all conditions. Only during the first three test cycles for *t* = 3 did the MSE remain constant. This is caused by the fact that in this condition the first three test cycles are linked nonadaptive tests. For $t = 3$ and $t = 5$ the MSE dropped at the $(t + 1)^{\text{th}}$ test cycle, which could be explained by the increasing adaptivity stemming from the conversion of the calibration cluster into an additional adaptive cluster. Up to the $(t + 1)^{\text{th}}$ test cycle the respective slower calibration procedure showed a lower MSE, especially for $N = 50$. For the ninth test cycle, the different levels of calibration speed seem to perform similarly for each sample size with a little disadvantage of the $t = 9$ condition, which could be explained by the fact that there is still a calibration cluster in the ninth test cycle and therefore less adaptivity in this condition only.

Over the course of the test cycles, the MSE converged to the lower baseline and moved away from the upper baseline in all conditions. However, for $N = 300$ there were nearly no differences between the MSE stemming from the continuous calibration and the upper and the lower baseline. For this sample size, using the initial estimation of the item parameters (without updating the parameters) works approximately as well as using the real item parameters. The increasing precision is apparently based only on the growth of the item pool.

Figure 3 shows how the MSE evolved over the test cycles under the 2PL model. The results resemble the results of the 1PL, with the exception that the lowest MSE was reached for $t = 9$ over the course of the nine test cycles. In addition, for $t = 3$, the MSE increased from test cycle one to test cycle two even for $N = 300$, which could be interpreted as an effect of the linking procedure.

Conditional precision

In a second step, the conditional precision of the ability estimates for specific ranges of the ability scale was investigated. Figure 4 illustrates the results for the 1PL. For the sake of clarity of presentation, the results are only presented for the first, the third, the fifth and the ninth test cycles. As seen, in every condition and in all test cycles in the calibration process, the precision of ability estimates was highest for medium level ability scores. Over the course of the test cycles, the difference in measurement precision between test takers with extreme ability scores and test takers with medium ability scores decreased. For t = 3, the results of the first and the third test cycles, similar to the results of the fifth and the ninth test cycles, were fairly equal. Updating the item parameters in this condition had only very small impact on the conditional precision.

Figure 5 illustrates the results under the 2PL model. These were very similar to those obtained under the 1PL model, but for $t = 3$, similar to global precision, the conditional precision decreased from test cycle one to three. Between $t = 5$ and $t = 9$ there were nearly no differences in the conditional precision over the course of the test cycles.

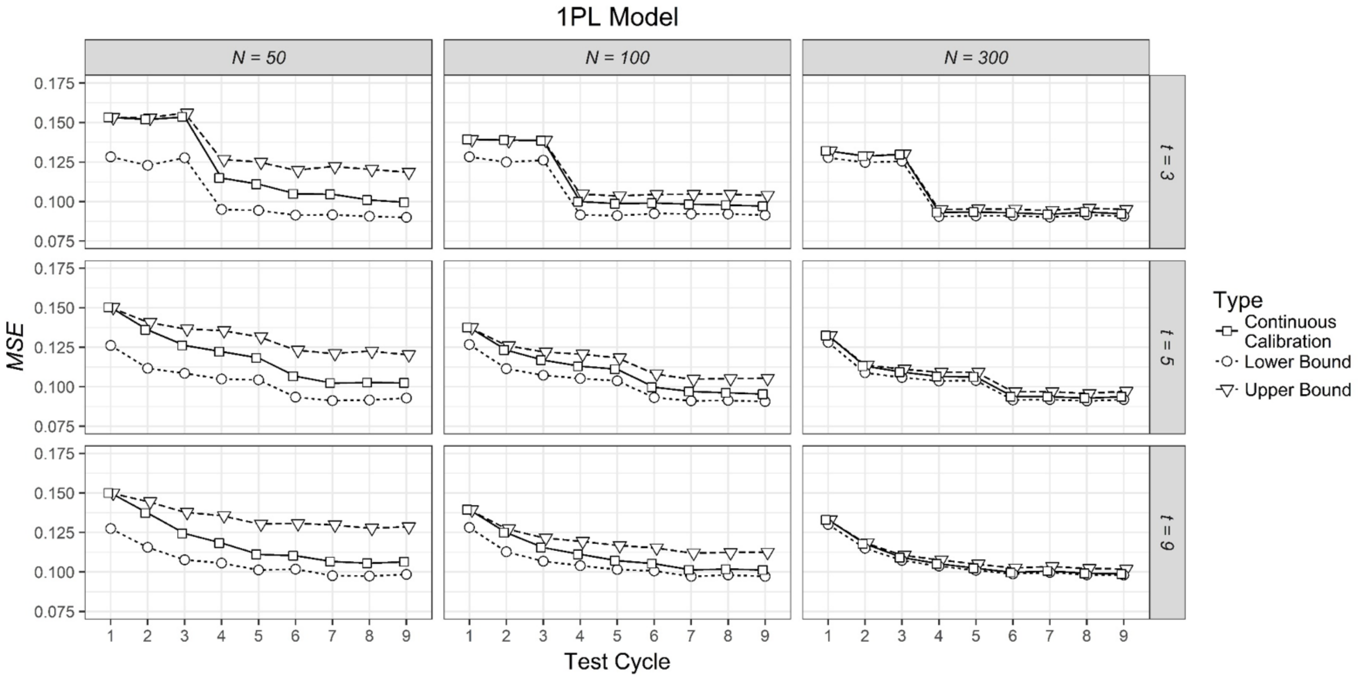

Figure 2:
Global mean squared error (*MSE*) for each test cycle (*t*) in the continuous calibration strategy for different sample sizes and different levels of calibration speed under the 1PL model.

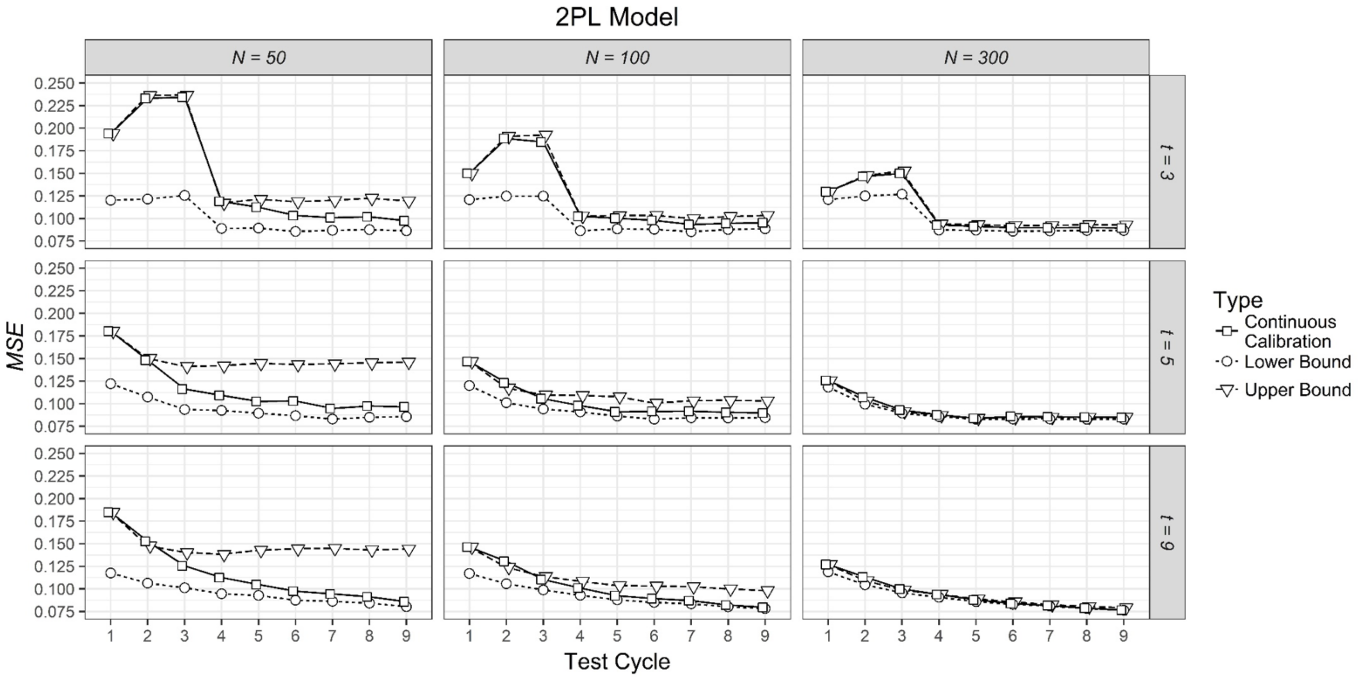

Figure 3:
Global mean squared error (*MSE*) for each test cycle (*t*) in the continuous calibration strategy for different sample sizes and different levels of calibration speed under the 2PL model.

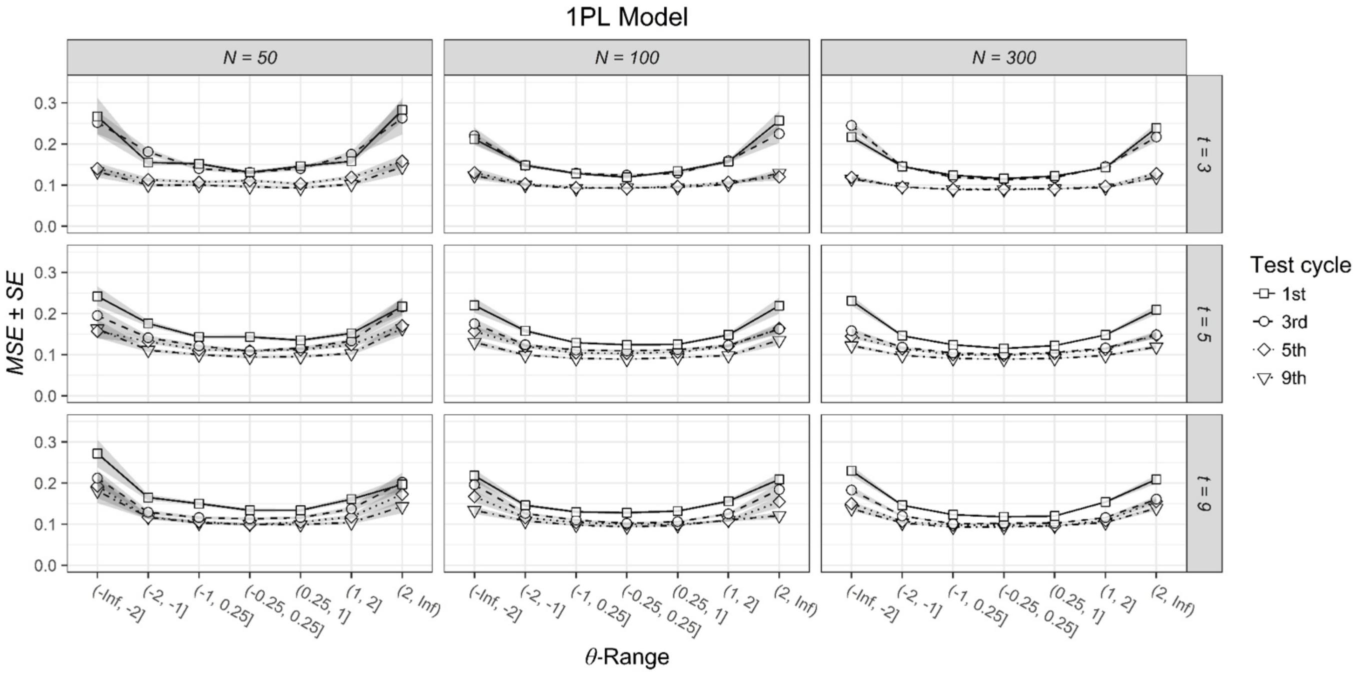

Figure 4:
Conditional mean squared error (*MSE*) at specific areas on the θ-continuum after different numbers of test cycles in the continuous calibration strategy for different sample sizes and different levels of calibration speed under the 1PL model. The areas shaded gray represent the standard error around the calculated MSE obtained from the variability of all simulees in a specific θ-range.

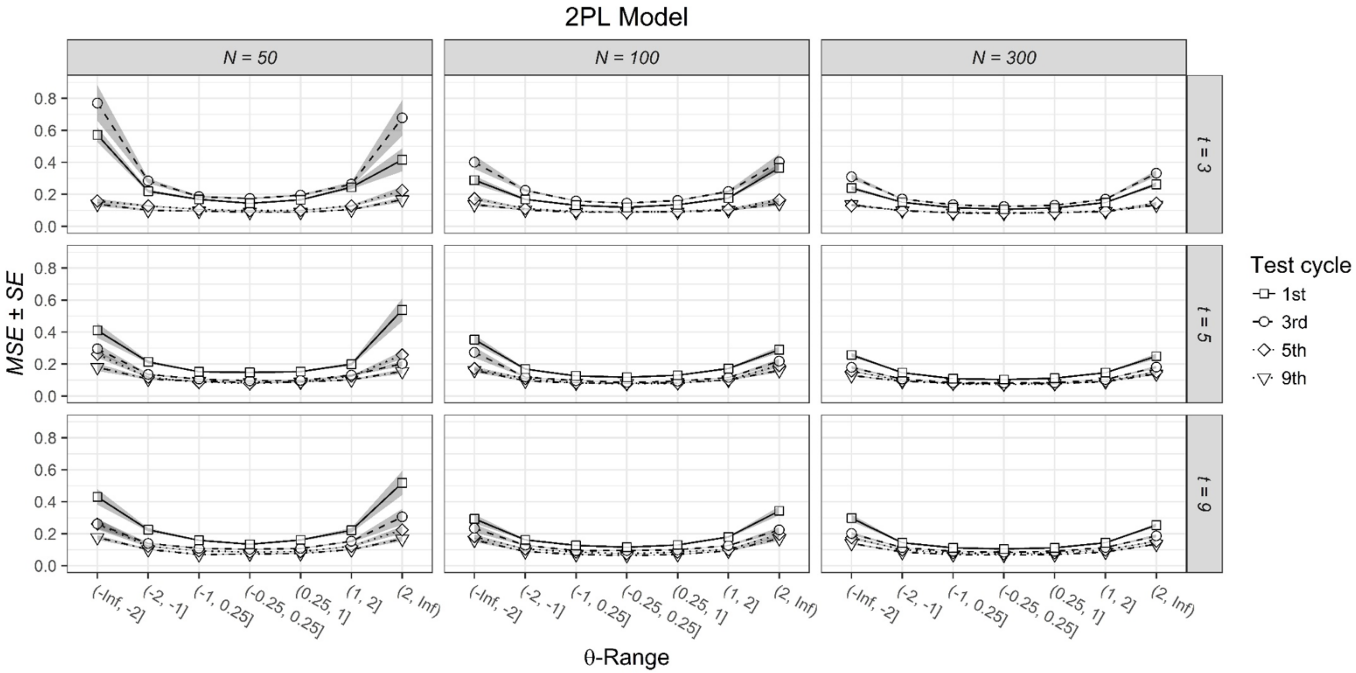

Figure 5:
Conditional mean squared error (*MSE*) at specific areas on the θ-continuum after different numbers of test cycles in the continuous calibration strategy for different sample sizes and different levels of calibration speed under the 2PL model. The area shaded gray represents the standard error around the calculated MSE obtained from the variability of all simulees in a specific θ-range.

Discussion

The general purpose of this study was to propose and examine a new continuous calibration strategy for the application of CAT in areas, where it is infeasible or difficult to conduct a calibration study before the operational phase of a computerized adaptive test and it is tolerable that measurement precision and adaptivity increase across test cycles. Such a method can be used immediately during everyday testing practice without the need for a separate calibration study. The basic ideas of the strategy are (a) constantly updating item parameters across several test cycles by using all the information from preceding test cycles, (b) adaptive item administration of calibrated items, (c) online-calibration of new items, and (d) maintaining the scale across several test cycles through linking and, therefore, allowing comparisons of the ability measures across these test cycles.

To give some practical recommendations for the configuration of the continuous calibration strategy, a simulation study was carried out, and basic tuning parameters were varied. These were (1) the sample size per test cycle, (2) the level of calibration speed, and (3) the underlying IRT model. Regarding the sample size, the results show a promising performance of the proposed strategy even for very small samples of $N = 50$. For $N = 300$ there is nearly no difference between the upper and lower baselines and the precision obtained with the continuous calibration strategy. Thus, for sample sizes such as this, it is not necessary to update the item parameters over test cycles. Nonetheless, also in this case, a test for IPD should be implemented.

With respect to the calibration speed, when using the 1PL, there is no clear recommendation. Test developers have to weigh the costs and the benefits of either a slow or a fast initial calibration. If it is tolerable that there is a substantial increase in measurement precision and if there are enough resources to construct many items for the first few test cycles, it is certainly possible to calibrate many items as fast as possible. In contrast, for the 2PL model, a slow calibration clearly outperformed a fast calibration regarding global and conditional measurement precision and therefore is recommended especially for very small samples of $N = 50$. The differences between the $t = 5$ and the $t = 9$ conditions were very small. Thus, depending on the resources available for producing new items, test developers should decide between a medium and a low level of calibration speed. A fast calibration under the 2PL model, especially in small samples, is not recommended. The selection algorithm in the 2PL model tends to select items with high discrimination parameters, which is not desired at early stages of the continuous calibration process because of larger calibration errors. For example, a good item could receive a small discrimination parameter by chance due to only a few inconsistent responses in a small sample. Therefore, this item might never get updated sufficiently, as it is simply not administered sufficiently often. As shown by van der Linden and Glas (2000), capitalization on calibration errors strongly impacts the ability estimation using the 2PL model. Thus, enlarging the item pool slowly leads to a more uniform item administration, as in the first few test cycles there are only few items the selection algorithm can choose from, which in turn leads to a higher number of item responses and more precise ability estimation.

In addition to the factors considered in this study, the proposed strategy has several interesting operational characteristics and tuning parameters that future research should address to investigate the performance of the proposed strategy in more detail and give practical recommendations. Some of them are, for example, the test length, the method used to test for IPD, the scale transformation method, the link item selection method, and the proportion of link items. Additionally, this study was carried out under the assumption that all items follow the respective model. Future research might also investigate the impact of item misfit on the performance of the continuous calibration strategy. Therefore, methods for assessing fit in items during the calibration process should be implemented. Finally, the implemented cluster structure provides the possibility of considering item position effects during the calibration process (e.g., Frey, Bernhardt, & Born, 2017). For this purpose, the three item clusters could be divided into more fine-grained, equal-sized subclusters and presented in a balanced way across different positions.

The major conclusion that can be drawn from the results is that the continuous calibration strategy works reasonably well, even for very small samples of N = 50. The proposed strategy offers a practical and less resource-consuming method for test developers from application areas with the characteristics mentioned above to take advantage of the benefits of CAT.

Acknowledgements

The research reported in the article was supported by a grant from the German Federal Ministry of Education and Research (Ref: 16DHL1005).

References

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28,* 147–162. doi: 10.1111/j.1745-3984.1991.tb00350.x

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika*, *46*, 443–459. doi: 10.1007/BF02293801

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275–285. doi: 10.1111/j.1745–3984.1988.tb00308.x

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48(6),* 1–29. doi:10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71(5),* 1–39. doi:10.18637/jss.v071.i05

Chen, P. (2017). A comparative study of online calibration methods in multidimensional computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 42,* 559–590. doi: 10.3102/1076998617695098 de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford.

Dolan, R. P., & Burling, K. S. (2012). Computer-based testing in higher education. In C. Secolsky & D. B. Denison (Eds.), *Handbook on measurement, assessment, and evaluation in higher education* (pp. 312–335). New York, NY: Routledge. doi: 10.4324/9780203142189.ch22

Frey, A. (2012). Adaptives Testen [Adaptive testing]. In H. Moosbrugger & A. Kelava (Eds.), Testtheorie und Fragebogenkonstruktion (2nd ed., pp. 275–293). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-20072-4_11

Frey, A., Bernhardt, R., & Born, S. (2017). Umgang mit Itempositionseffekten bei der Entwicklung computerisierter adaptiver Tests [Accounting for item position effects in the development of computerized adaptive tests]. *Diagnostica*, *63*, 167–178. doi: 10.1026/0012-1924/a000173

Frey, A., & Ehmke, T. (2007). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards [Hypothetical usage of adaptive testing for the examination of educational standards]. *Zeitschrift für Erziehungswissenschaft*, *Sonderheft 8*, 169–184. doi: 10.1007/978-3-531-90865-6_10

Goldstein, H. (1983). Measuring changes in educational attainment over time: problems and possibilities. *Journal of Educational Measurement, 20*, 369–377. doi: 10.1111/j.1745-3984.1983.tb00214.x

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22,* 144–149. doi: 10.4992/psycholres1954.22.144

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26,* 3–24. doi: 10.1177/0146621602026001001

He, W., & Reckase, M. D. (2014). Item pool design for an operational variable-length computerized adaptive test. *Educational and Psychological Measurement*, *74,* 473–494. doi: 10.1177/0013164413509629

Kingsbury, G. G. (2009). Adaptive item calibration: A simple process for estimating item parameters within a computerized adaptive test. *GMAC conference on computerized adaptive testing*. Minneapolis, MN.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer. doi: 10.1007/978-1-4939-0317-7_10

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*, 179–193. doi: 10.1111/j.1745-3984.1980.tb00825.x

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

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139–160. doi: 10.1111/j.1745-3984.1977.tb00033.x

Park, Y. S., Lee, Y.-S., & Xing, K. (2016). Investigating the impact of item parameter drift for item response theory models with mixture distributions. *Frontiers in Psychology, 255*(7), 1–17. doi: 10.3389/fpsyg.2016.00255

Patz, R. J., & Junker B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods in item response models. *Journal of educational and behavioral Statistics*, *24*, 146–178. doi: 10.2307/1165199

R Core Team (2017). *R: A language and environment for statistical computing* [Software]. R Foundation for Statistical Computing. Available from www.r-project.org

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 429–438). Boston: Elsevier Academic. doi: 10.1016/b0-12-369398-5/00444-8

Spoden, C., Frey, A. & Bernhardt, R. (in press). Running a CAT development within 18 months. *Journal of Computerized Adaptive Testing*.

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29,* 6–26. doi: 10.1016/S0361-476X(02)00063-2

Stocking, M. L. (1988). Scale drift in online calibration. *ETS Research Report Series 1988*(1), 1–122. doi:10.1002/j.2330-8516.1988.tb00284.x

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210. doi:10.1177/014662168300700208

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, *16(1).* Retrieved from www.pareonline.net/getvn.asp?v=16&n=1.

van der Linden, W. J. (2016). *Handbook of item response theory, volume one: models*. London: Chapman and Hall.

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: Springer. doi: 10.1007/978-0-387-85461-8

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450. doi: 10.1007/bf02294627

Wells, C. S., Hambleton, R. K., Kirkpatrick, R., & Meng, Y. (2014). An examination of two procedures for identifying consequential item parameter drift. *Applied Measurement in Education, 27*, 214–231. doi: 10.1080/08957347.2014.905786

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26*, 77–87. doi: 10.1177/0146621602261005

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347–364. doi: 10.1177/014662168400800312

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1–17. doi: 10.1207/s15326977ea1001_1