Journal of Computerized Adaptive Testing

Volume 6 Number 3

September 2018

Applications and Implementations of CAT

Implementing Three CATs Within Eighteen Months

Christian Spoden, Andreas Frey, and Raphael Bernhardt

DOI 10.7333/1809-060338

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing
www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing

Volume 6, Number 3, September 2018
DOI 10.7333/1809-060338
ISSN 2165-6592

Applications and Implementations of CAT

Implementing Three CATs Within Eighteen Months

Christian Spoden
Friedrich Schiller University Jena, Germany

Andreas Frey
Friedrich Schiller University Jena, Germany
and Centre for Educational Measurement (CEMO)
at the University of Oslo, Norway

Raphael Bernhardt
Friedrich Schiller University Jena, Germany

The development of a computerized adaptive test is considered a labor-intensive and time-consuming endeavor. This paper illustrates that this does not have to be the case—by demonstrating the steps taken, the decisions made, and the empirical results obtained during the development of three computerized adaptive tests (CATs) designed to measure student competencies in reading, mathematics, and science. The three tests had to be developed and piloted within an 18-month period, and they were used directly afterward in six research projects of a large nationwide research initiative. To ensure the sound psychometric quality of the CATs developed, the item calibration (N = 1,632) followed several quality control procedures, including item fit analysis, differential item functioning analysis, and preoperational simulation studies. A CAT pilot study (N = 1,093) and an additional CAT simulation confirmed the general usefulness of the constructed instruments. It is concluded that the development of CATs—including item calibration, simulations, and piloting within 18 months—is quite possible, even for comparably small development teams. This necessitates an available theoretical framework for the assessment and a sufficient number of items, specific plans for the item calibration, simulations, and a pilot study, as well as an information technology infrastructure for administering the tests.

Keywords: computerized adaptive testing, item calibration, item fit analysis, differential item functioning, measurement precision

General competencies in areas such as reading, mathematics, and science describe how well individuals can apply the knowledge and skills that they have acquired during their education to real-life challenges. More specifically, general competencies are considered powerful predictors of academic achievement for students in vocational education and training (VET; e.g., Helm, 2014, 2015; Seeber & Lehmann, 2013); as such, they are important prerequisites for success in the labor market. Consequently, general competencies played an important role in a nationwide research initiative concerned with technology-based competence assessments for VET in Germany. This research initiative was funded by the German Federal Ministry of Education and Research [Bundesminsterium für Bildung und Forschung (BMBF)] and took place from 2011 to 2014. It included six joint projects that consisted of 21 individual projects. The main objective of all the projects was to construct computer-based tests to measure the competencies needed to be successful in diverse professions, such as technicians, industrial clerks, medical assistants, and caregivers for the elderly. For all profession-specific competencies measured, the relation with general competencies in reading, mathematics, and science was examined.

To avoid parallel work and to foster comparability among the individual projects, the tests for reading, mathematics, and science were centrally developed. Computerized adaptive testing (CAT) allowed for the development of three testing instruments that were examined in the research initiative and could be used with comparable precision across the professions. CAT was used for two major reasons. First, it makes it possible to efficiently measure competencies across a large ability range. Such a large range was expected in this sample because the minimum number of school years required in the professions examined varied from 8 to 12 years. Second, the testing time needed to measure the general competencies needed to be short so that the individual projects could allocate a greater amount of time to the development of their profession-specific tests.

The construction of the three CATs to measure reading, mathematics, and science had to be completed within the time interval of not more than 18 months because the tests were needed by the end of that time for the six research projects. Finalizing stable operational tests for the three general competencies by this deadline was crucial because delays would have caused problems for all six research projects. However, there was neither a benchmark nor any previous experience available on how to complete the test development process—which included item bank composition, item calibration, preoperational simulation studies, and field testing of the tests and the test software—in such a short time frame. Fulfilling this objective was not only a serious challenge but also a situation where the established routines of CAT development could not be adopted due to the tight time frame. Therefore, new solutions had to be found. The present paper exemplifies the decisions made as well as the work steps taken to establish a robust and user-friendly CAT system within 18 months. It focuses on item calibration, psychometric analysis by means of real and simulated data analysis, and a CAT pilot, demonstrating how an efficient CAT development process can be realized while simultaneously meeting high quality standards.

This paper is organized as follows. The next section describes the general reasons for implementing a user-friendly CAT for the assessment of general competencies in reading, mathematics, and science. The third section outlines analyses of the responses obtained in the calibration study that were conducted to configure the three CATs. This includes estimating item difficulties, analyzing item fit and differential item functioning (DIF), and conducting preoperational simulation studies. The fourth section illustrates a CAT pilot study that was conducted after the completion of the item bank and its implementation in an adaptive administration system, as well as a CAT simulation study conducted prior to test release. The timeline of CAT development is described in the fifth section. In the sixth section, conclusions

from the current test development process are presented, and recommendations for future time-critical CAT developments are offered.

General Decisions Made When Establishing a High-Quality CAT

Developing CATs for each of the three domains—reading, mathematics, and science—within 18 months was a major challenge, even for experienced test developers. Several general decisions had to be made before the items were calibrated concerning (1) the theoretical framework, (2) the mode of test delivery, and (3) the development of item banks. Thus, an efficient CAT implementation was only possible by taking several shortcuts compared to the conventional test development process (e.g., Lane, Raymond, Haladyna, & Downing, 2015).

Theoretical Framework

A clear definition of the content areas under study in terms of a theoretical framework is a prerequisite for successful test development. Thus, the main objective was to find a solution that provided the best possible quality in a minimum amount of time. It was therefore decided to adopt existing theoretical frameworks. International large-scale assessments such as the Programme for International Student Assessment (PISA; Organisation for Economic Cooperation and Development [OECD], 2016) and the Trends in International Mathematics and Science Study (TIMSS; Mullis & Martin, 2013) have provided frameworks for the assessment of general competencies based on a broad consensus among content-matter experts. For the present test development process, content-matter experts inspected such theoretical frameworks to determine applicability for an assessment of VET students; and, if deemed necessary, the content experts altered the frameworks accordingly. Finally, the test contents of the three CATs were defined as follows.

The theoretical framework for *reading competence* was based on the theoretical framework of PISA 2009, in which reading competence is defined as "understanding, using, reflecting on and engaging with written texts in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society" (OECD, 2009, p. 23). These skills should be demonstrated for functional reading texts, continuous texts, and discontinuous texts. Additionally, the theoretical framework for reading in the research initiative included three levels of cognitive activities—namely, *identification*, *integration*, and *generation*—derived from the model of learning from verbal and pictorial representations by Schnotz and Bannert (2003). The framework operationalized the principles of functional reading competence, as described in detail by Ziegler, Balkenhol, Keimes, and Rexing (2012). Compared to reading as it occurs in schools—where the primary goal is to learn new content—functional reading competence not only refers to a specific technical language and text that is specific to a vocational field but also to reading with the goal of completing authentic, real-world tasks. This means that there is a focus on different pieces of information in the text and on the production of the functional and structural analogue representations of the content of the text (Ziegler et al., 2012). The content-matter experts checked that the different tasks did not require too much domain-specific knowledge or technical terminology and that contexts and text genres were considered that were relevant, to some extent, for each of the different VET professions assessed in the research initiative. For this purpose, curricula for the different professions were reviewed before the development of the initial item banks for the reading test.

With respect to the conceptual basis for the assessment of *mathematical competence*, the theoretical framework of PISA 2009 (OECD, 2009) was adopted. The framework focuses on the engagement with mathematics and, hence, on using and doing mathematics in a variety

of situations to solve problems, including personal, educational/professional, public, and scientific situations. Additionally, the content is organized into four overarching ideas—namely, *space and shape*, *change and relationships*, *quantity,* and *uncertainty*. Within each of these overarching ideas, different levels of cognitive processing are distinguished and grouped into three competence clusters: *reproduction*, *connections*, and *reflection*.

The theoretical understanding of student *competence in science* is based on a combination of the theoretical science framework for Grade 8 of TIMSS (Garden & Orpwood, 1996) and the framework of the Swiss national educational standards for science (Konsortium HarmoS Naturwissenschaften+, 2009). Four content strands—namely, *biology*, *earth science, chemistry*, and *physics*—are specified as subdomains. Additionally, three levels of cognitive processing are differentiated. These are *comprehension of simple and complex information*, *analyzing and conceptualizing*, and *applying scientific evidence to decisions in complex situations*.

Mode of Test Delivery

In order to keep the costs low while ensuring a fast implementation process, an existing noncommercial software package had to be used for the delivery of the three new adaptive tests. After the requirements of the six research projects were identified, it was decided to use the Multidimensional Adaptive Testing Environment (MATE; Kröhne & Frey, 2013) for both the calibration study and the field test. A web-based delivery approach was the most efficient approach for the six projects. The adaptive tests developed were delivered via the internet with a secure connection, and the responses were stored on a central university server. The computers used for testing required a connection to the internet and the installation of MATE and Microsoft Silverlight. To avoid delays caused by a low bandwidth internet connection for tests in group settings, the item banks were downloaded prior to the testing day. In this way, only the responses and the information concerning which item should be presented next were transferred between the testing clients and the central server during testing.

Standardized testing situations were used. In all participating schools, testing took place in group settings using existing school hardware. The monitors had a minimum screen size of 15 inches. The MATE software adjusts the presentation of each test item to the screen size to prevent degrading text or graphic images. Prior to the testing date, the local test supervisors also checked that all test items were readable and that the items were presented without any delay. Differences in the computational power between computers were of low importance, as no applications with heavy computational demands (such as video streaming) were included and the computations relevant for the CAT functioning were performed by the central server, which had sufficient computational power.

Development of the Initial Item Banks

To save time, one goal of the development of the initial item banks was to write as few new items as possible. Rather than writing new items, items that had already been constructed and field tested were preferred for the item banks. As item sources, published items available in German from international and national large-scale assessments of student achievement and items that had been field tested but not yet used in operational tests (e.g., because there already were enough items in the same difficulty range) were selected. Considering the population of VET students in the research initiative, items were adapted to better fit the vocational context, if necessary. All selected items were originally used in paper-and-pencil format. Therefore, the next task was to computerize the items. To facilitate the computerization, only items that could be scored directly during an adaptive test were selected (multiple-choice and short-answer response format). Content-matter experts from each of

the three domains were asked to evaluate each item with respect to the following criteria: (1) fit to the respective theoretical framework, (2) formal correctness (including layout and completeness in terms of item stimuli and item options), (3) scoring correctness, (4) appropriateness of the item for the VET student population, and (5) readability of the visual presentation of the item on a computer screen.

Based on these criteria, initial item banks were assembled, covering all aspects of the theoretical frameworks (e.g., different cognitive processes or strands) in a balanced way, as well as covering a broad ability range (Table 1). These item banks were then used in the calibration study. They comprised 73 items for reading from three subdomains (functional reading texts, continuous texts, and discontinuous texts); 133 items for mathematics from four subdomains (space and shape, change and relationships, quantity, and uncertainty); and 133 items for science from four subdomains (biology, earth science, chemistry, and physics). Note that items were developed according to these specified subdomains to ensure that conclusions about the competence of students were drawn based on item sets that adequately represented the width of those domains in terms of content.

With respect to reading, approximately half of the items were rewritten because not enough field-tested items could be found for some aspects of the reading framework. The smaller size of the reading item bank compared to the other two item banks also reflects, to a certain degree, that the reading items take approximately twice as long to answer. The item banks for mathematics and science consisted completely of items stemming from other assessments. The copyright holders of all existing items selected granted permission to use their items in the new CATs.

Although relying predominantly on existing items, the development of the initial item banks still took approximately five months. This indicates that meeting the time limit of 18 months was only possible by selecting items from available and tested item banks.

Table 1. Number of Items per Domain and Subdomain in the Initial Item Banks and Number of Items Eliminated From the Item Bank Due to Item Misfit or Low Levels of Item Discrimination, and DIF

Domain and Subdomain	Original Item Bank Size	Eliminated Items		Reduced Item Bank Size
		Item Fit and Discrimination Analysis	DIF Analysis	
Reading				
Functional Reading Text	28	1	1	26
Continuous Text	26	2	2	23
Discontinuous Text	19	0	0	19
Mathematics				
Space and Shape	34	4	1	30
Change and Relationships	32	3	1	28
Quantity	36	6	1	30
Uncertainty	31	5	3	23
Science				
Biology	33	5	1	27
Earth Science	33	7	1	22
Chemistry	35	5	6	27
Physics	32	8	3	20

Note. Items may be flagged due to item misfit or low item discrimination, and differential item functioning.

Calibration Study

The items from the initial item banks were trialed in a computer-based, nonadaptive item calibration study. The major aim of this calibration study was to derive precise item parameter estimates for the operational CAT phase. Additional aims were to ensure the fit of the items with the item response theory (IRT) model used across several subpopulations of VET students and to examine whether precise ability estimates could be expected from future CAT administrations if the item selection is restricted to the present items for all subdomains, as specified in the theoretical frameworks. To accomplish these aims, the analyses connected with the calibration study included the estimation of item parameters, an item fit analysis, a DIF analysis, and a preoperational simulation study.

Student Sample and Data Collection

The first three steps were conducted based on item calibration data collected at 27 vocational schools in the German federal states of Hesse, Lower Saxony, and Thuringia. The data collection took place at schools because VET is organized within a dual education system in German-speaking countries. This system combines apprenticeships in a company with vocational education at a vocational school. The majority of the programs in the dual education system have a duration of three years.

The sample tested consisted of N = 1,632 VET students (46% female; 87% with German as primary language) from commercial and administrative professions (N = 566), technical or industrial professions (N = 658), or nursing or other professions in the medical field (N = 408). The mean age in the sample was M = 21.38 years (SD = 3.03). The majority of students (68%) were in their third year of vocational training.

Each tested student had to work on one test composition (a specific test form containing a subset of the complete item bank) based on a complex test design. As the number of items to be calibrated was too large to present each item to every student, a balanced incomplete block design with two levels was used to assemble the test compositions. At Level 1, a Youden square design (e.g., Giesbrecht & Gumpertz, 2004; Preece, 1996) was used for each content area (reading, mathematics, and science). Thereby, for each domain, it was ensured that across all test compositions (1) all tests had the same length, (2) all items were presented with equal frequency, (3) each pair of two items was presented with equal frequency, and (4) each item was presented in every possible position with equal frequency. Thus, possible item position effects (e.g., Debeer & Janssen, 2013; Nagy, Nagengast, Frey, Becker, & Rose, 2018) that might be caused by growing fatigue or a decline in motivation toward the end of the test were balanced across test items (Frey, Hartig, & Rupp, 2009). The domain-specific designs from Level 1 were nested in the second level.

At the second level, three blocks were specified into which items from one content area were grouped. Therefore, each test composition contained one block of reading items, one block of mathematics items, and one block of science items. To balance potential order effects, all possible block orders were specified by complete permutation. Thereby, the structure at Level 2 ensured that (5) each test composition contained one block for each content area, and (6) each possible ordering of the three blocks was used with equal frequency across the set in order to test compositions. Each test composition consisted of 33 items—nine items from reading and 12 items each from mathematics and science. The testing time allotted was 40 minutes. As the students needed an average of 21 minutes (SD = 9) to complete the test, the test was considered as a power test. Missing item scores in the data occurred when items were either not reached or not administered to a student due to the balanced incomplete block design; these item responses were treated as missing completely at random in the IRT

scaling. The design resulted in about 200 valid responses for reading items and about 150 valid responses each for mathematics and science.

The tests were browser-based and the data were transferred to and stored on a central university server after each response. The schools participated voluntarily and received a summary of the results after the test.

The complete procedure for the data acquisition and data analysis was planned beforehand and required approximately two months of working time. The schools were contacted well in advance of the testing dates, and a date was selected for the administration of the test. The code to be executed for data handling and data analysis was prepared to as great a degree as possible, and the item selection criteria were fixed a priori. Thus, it was possible to acquire the data and process the calibration data quickly and efficiently in a period of approximately two additional months.

Item Calibration

The basic aim of the item calibration study was to derive precise item parameter estimates for the operational CAT phase based on an IRT model (e.g., van der Linden, 2016). Item calibration was based on the unidimensional Rasch model (Rasch, 1960/1980) given by

$$\Pr(x = 1 | \theta_u, \delta_i) = \text{logit}(\theta_u - \delta_i) \quad [1]$$

with

$$\text{logit}(y) = \frac{e^y}{1 + e^y}, \quad [2]$$

where θ_u denotes the ability of student u and δ_i denotes the difficulty of item i. The Rasch model was chosen because the sample size requirements are smaller for this model than for IRT models with more parameters (e.g., de Ayala, 2009). Additionally, the psychometric checks to establish the invariance of item parameters across gender and the professions examined, conducted with DIF analyses, can be carried out faster. Finally, the specification and interpretation of the θ scale referring to the preserved order of items throughout the θ range are straightforward with this model (Wilson, 2003), although this is not the case for the more complex IRT models including those with a discrimination parameter. The estimation of item parameters, as well as the following item fit and DIF analyses, were carried out with ConQuest 3.0 (Adams, Wu, & Wilson, 2012) using marginal maximum likelihood estimation (MLE). For identification purposes, the mean of the θ distribution was set to zero (with SD not constrained) and all item slopes were set to 1.

Item Fit Analyses

The next step in the analyses was the item fit analysis to ensure adequate conformity to the Rasch model in the presence of items with diverse content within each domain. This item fit analysis was based on the weighted mean square error (WMNSQ; Wright, 1980) of the item difficulty parameters and a critical value of t(WMNSQ) > 1.96, as well as an empirical item discrimination obtained by the item-total correlation of $r_{it} > .25$. Table 1 provides an overview of the number of items eliminated from the three item banks due to item misfit or low item discrimination, and DIF (see discussion below). The results presented in this table show that a rather small number of items were eliminated according to these criteria, a finding that can be attributed to the fact that all items were thoroughly checked by content-matter

experts before administration and that the majority of the items had previously been field tested.

Differential Item Functioning Analyses

As the third step, DIF (Holland & Wainer, 1993) was examined. Item calibration datasets often include a larger amount of background information on the participants and thus a substantial number of possible DIF variables. To ensure a focused and efficient analysis of relevant DIF variables, gender and the profession of the students were examined as the two covariates of special importance. Gender DIF was analyzed, as previous large-scale assessments in reading and mathematics found substantial gender differences in Germany (e.g., Mullis, Martin, Foy, & Arora, 2012; OECD, 2014). This not only stressed the general impact of gender in the German school system but also raised the question concerning the extent to which its impact can be explained by the bias of the test items. A DIF analysis between students from different professions was conducted to guarantee the usefulness and applicability of the testing instruments across different VET professions so that the test scores derived could be compared across professions. The DIF analysis was conducted in a model-based way according to the multifacet Rasch model (MFRM; Linacre, 1989). Thus,

$$\mathrm{P}\left(x = 1 | \, \theta_u, \delta_i, \lambda_g, \xi_u\right) = \mathrm{logit}(\theta_u - \delta_i - \lambda_g - \delta_i\lambda_g - \xi_u) \, , \qquad [3]$$

where θ_u and δ_i are as defined above, λ_g denotes the mean ability of the *g*th group of interest in the DIF analysis, $\delta_i\lambda_g$ denotes the interaction term of item difficulty and group membership, and ξ_u denotes potential additional covariates to be controlled for in a DIF analysis. Use of the MFRM for DIF analyses is described, for example, in Engelhard (2009). For details on DIF analyses using the ConQuest software, which allows the expression of the DIF effect size in terms of a *t*-value, see Wu, Wilson, Adams, and Haldane (2007, p. 80-90).

Items exhibiting significant DIF ($t > 1.96$) were not directly eliminated from the respective item bank but were additionally checked by content-matter experts. The items were eliminated if the content-matter experts concluded that the respective DIF variable (gender or profession) was likely to induce construct-irrelevant variance, and they were retained if this was not the case. In addition to the results on item fit and empirical discrimination, Table 1 also provides an overview of the number of items eliminated when both criteria—statistical information on DIF and judgments of content-matter experts—indicated item bias (see Zumbo, 1999). A more detailed DIF analysis for the present data examining domain- and item-specific effects between commercial and administrative jobs, and technical and industrial jobs, is provided in Spoden et al. (2015).

During the course of the item fit analysis and the DIF analysis, the items were again checked for scoring problems and problems with item distractors; this resulted in the elimination of one additional item for reading and 11 items for science. The calibrated item banks for the three computerized adaptive tests comprised 68 reading items, 111 mathematics items, and 96 science items. The item difficulty distributions revealed reasonable levels of variance in each of the three domains—reading (M = –0.38, SD = 1.12); mathematics (M = –0.28, SD = 1.32); and science (M = –0.720, SD = 1.10)—that were similar to the estimated θ ($\hat{\theta}$) distributions (reading, M = 0, SD = .86; mathematics, M = 0, SD = .98; science, M = 0, SD = .87). The means and variances of the $\hat{\theta}$ distributions were later applied as priors for θ estimation in each of the following steps of the test development.

Preoperational Monte-Carlo Simulations

The preoperational monte-carlo simulations were carried out with SAS 9.3 (SAS Institute Inc., 2011) and had two objectives. The first objective was to predict the precision of the θ estimates for each of the three adaptive tests and to compare SAS with nonadaptive testing. The result not only makes it possible to gauge the success of the development process but it also documents the advantages of using CAT in the present study, especially given that CAT was found to be advantageous across all six research projects. The second objective was to investigate the number of items selected per subdomain and to check whether these numbers met the required numbers according to the theoretical frameworks (see below).

The simulations used the calibrated item banks after item elimination. The CAT condition was contrasted with a nonadaptive fixed-item testing (FIT) condition. In order to control the proportions of items presented for the subdomains, and thereby to ensure that the content-related size of the calibrated item banks was maintained in the set of items administered to each student, the maximum priority index method (MPI; Cheng & Chang, 2009) was used. The MPI is a heuristic constraint management approach (Born & Frey, 2017). Within each step of the item selection, the priority index (PI) for each candidate item i^* from the respective calibrated item bank was computed, and the item with the highest PI was selected for administration. The PI for item i^* was computed according to

$$\mathrm{PI}_{i^*} = I_{i^*} \prod_{k=1}^{K} (w_k f_k)^{c_{i^*k}}, \qquad [4]$$

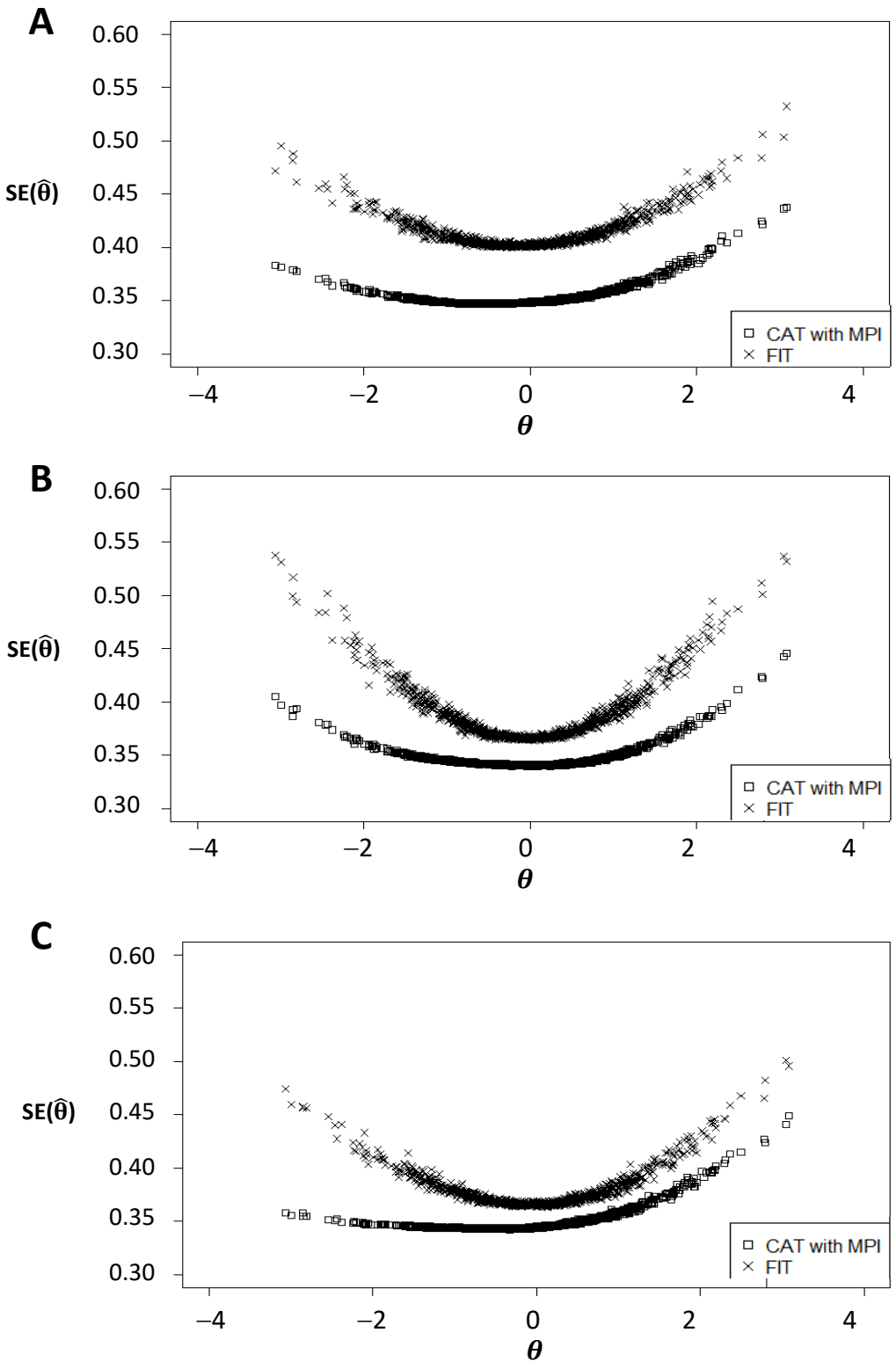
where I_{i^*} is the Fisher information based on the provisional $\hat{\theta}$, w_k is the weight of constraint k controlling the relative importance of the various constraints, and c_{i^*k} is a binary value indicating whether item i^* is relevant for the constraint k (c_{i^*k}= 1) or not (c_{i^*k}= 0). The term f_k is the quota left according to Cheng and Chang (2009) and is defined by

$$f_k = \frac{b_k - x_k}{b_k}, \qquad [5]$$

where b_k denotes the number of items required for the test that are relevant for constraint k, and x_k denotes the number of items administered that are relevant for constraint k. In the present study, equal proportions for each of the subdomains described above were targeted. Thus, for each subdomain the same values were used for b_k and all weights were set to 1.

To quantify the gains in measurement precision that can be expected by using CAT instead of FIT, the three CAT versions were compared with corresponding FIT versions in terms of the standard error of the θ estimate, SE($\hat{\theta}$), and the number of violations of the target proportions of items per subdomain. The simulations were conducted with 50 replications and a sample size of N = 1,000, based on the assumption of $\theta \sim$ N(0, 1). To make it possible to directly compare the results between the CAT and FIT conditions, the same number of items were administered in both conditions (32 items for reading, 36 each for mathematics and science). These test lengths were based on previous simulations showing that at least 30 items were required to achieve a reliability of .70 in the FIT condition. In the CAT condition, the first item was randomly selected from 10 items of medium difficulty, and the subsequent items were selected using the MPI, as specified above. Bayes modal estimation (BME) was used to

Figure 1. Standard Error SE($\widehat{\theta}$) as a Function of θ for FIT and CAT with MPI for Reading (A), Mathematics (B), and Science (C)



estimate θ and the SE was computed as the square root of the reciprocal value of the test information function at $\hat{\theta}$.

Figure 1 presents the results of the SE($\hat{\theta}$) for the three domains. As intended when applying computerized adaptive tests, the results from this simulation show a higher measurement precision for CAT with the MPI compared to FIT, especially for extreme θ levels. This holds, in particular, for the reading and mathematics tests, where the CAT version with the MPI substantially increased the measurement precision, as evident from an obviously lower SE($\hat{\theta}$) level. With respect to the science test, the same is apparent for lower θ levels, whereas the differences between CAT and FIT decline for higher θ levels, most likely as a consequence of a shortage of difficult items in the bank.

Additionally, checks were performed on how often the required number of items per subdimension was not met. No violations were observed in the sample of 50,000 simulated response vectors (1,000 participants × 50 replications), indicating that the item banks were large and diverse enough to represent the constructs to be measured according to the theoretical frameworks.

CAT Pilot Study and CAT Simulation Study

The calibrated item banks were used to assemble three adaptive tests that were then implemented using the MATE software (Kröhne & Frey, 2013). The three adaptive tests were used in a CAT pilot study and an additional CAT simulation study. The pilot study was conducted to ensure that the final CAT version worked as expected in practice. The subsequent monte-carlo simulation was carried out to predict target reliabilities for different test lengths. This enabled the six projects in the research initiative to assign the testing time necessary to measure competencies in reading, mathematics, and science.

CAT Pilot Study

A CAT pilot study with *N* = 1,093 students (38% female; 86% with German as the native language) from commercial or administrative professions (*N* = 374), technical or industrial professions (*N* = 515), nursing or other professions in the medical field (*N* = 173), and other or not specified professions (*N* = 31) was conducted within four months. The participants were mainly (71%) sampled from students who were in their third year of vocational training. The mean age in the sample was M = 22.06 years (SD = 3.74). Each of the students completed one of the three CATs in the domains of mathematics (*N* = 390), science (*N* = 353), or reading (*N* = 350). The particular domain administered to a student was randomly selected. Note that the overall sample size of students was larger than what would usually be required to pilot a CAT administration, as this study was also used for additional research on the motivational and affective effects of CAT versus FIT.

The following CAT algorithm was used in the CAT pilot study: The first item was randomly selected from 10 items of a medium level of difficulty, and the subsequent items were selected according to maximum information with the MPI to control the proportions of presented items per subdomain. BMEs were used as θ estimates. Two stopping rules of the CAT procedure were implemented to meet the requirements of the second objective previously outlined: the completion of 40 minutes of testing time or the administration of 48 items. In the CAT pilot study, two issues that were examined were whether the algorithm worked as previously simulated, that is, whether the variance of the $\hat{\theta}$ levels and the reliabilities were reasonably high, and whether the MPI selected items with equal probabilities from the several

subdomains of mathematics, science, and reading. Reliability was estimated from the empirical data by

$$\rho(\hat{\theta},\theta)^2 = \frac{\sigma(\hat{\theta})^2}{\sigma(\hat{\theta})^2 + \varepsilon\left(\sigma(\hat{\theta}|\theta)^2\right)} = \frac{\sigma(\hat{\theta})^2}{\sigma(\hat{\theta})^2 + \frac{1}{N}\sum_{u=1}^{N} \mathrm{SE}(\hat{\theta}_u)^2}, \quad [6]$$

where $\hat{\theta}$ is the BME of student u (Kim, 2012).

The statistical analysis of the responses from the pilot study revealed $\hat{\theta}$ distributions with reasonable levels of variance in each of the three domains—reading (M = –0.12, SD = 0.92); mathematics (M = –0.08, SD = 0.92); and science (M = –0.01, SD = 0.77)—although especially the variance in the science competence test was smaller than in the item calibration study. It was also found that the maximum number of items was not reached due to the time limit of 40 minutes. In the testing time allotted, the students answered M = 35.51 items (SD = 12.90) on average. Given the mean computed time on task for each item (i.e., reading, 100 seconds per item; mathematics, 64.9 seconds per item; science, 50.0 seconds per item), mean empirical reliabilities were computed for a fixed number of administered items in each of the three domains. The testing time was computed as the product of the test length and the mean time on task.

The results from these computations indicate that for reading, 12 items (four items in each of the three subdomains), on average, had to be answered to achieve a reliability estimate of $\rho(\hat{\theta},\theta)^2 > .70$ in 20 minutes; and 21 items (seven items in each of the three subdomains), on average, had to be answered to attain a reliability estimate of $\rho(\hat{\theta},\theta)^2 > .80$ in 35 minutes. For mathematics, 12 items (three items in each of the four subdomains), on average, had to be answered to achieve a reliability estimate of $\rho(\hat{\theta},\theta)^2 > .70$ in 13 minutes; and 20 items (five items in each of the four subdomains), on average, had to be answered to reach a reliability estimate of $\rho(\hat{\theta},\theta)^2 > .80$ in 22 minutes. Finally, for science, 16 items (four items in each of the four subdomains), on average, had to be answered to achieve a reliability estimate of $\rho(\hat{\theta},\theta)^2 > .70$ in 13 minutes; and 28 items (seven items in each of the four subdomains), on average, had to be answered to attain a reliability estimate of $\rho(\hat{\theta},\theta)^2 > .80$ in 23 minutes. These reliabilities underscore a high level of measurement precision within a reasonable testing time.

Table 2. Percentage of Items Selected by the MPI Content Balancing Procedure Across All Participating Students for Three Reading Subdomains, Four Mathematics Subdomains, and Four Science Subdomains (Adapted from Bernhardt, 2017)

Domain	Subdomain 1	Subdomain 2	Subdomain 3	Subdomain 4
Reading	33.63%	33.81%	32.56%	--
Mathematics	24.92%	25.34%	25.27%	24.46%
Science	24.94%	25.08%	25.17%	24.81%

Table 2 presents the results for the percentage of items selected per subdomain, using the MPI, across all participating students for reading, mathematics, and science. The percentages are close to the theoretical values of 33.33% (reading) and 25% (each for mathematics and science) per subdomain. The small deviations from the theoretical values are due to the stopping rules imposed so that equal proportions of presented items per subdomain were not possible in every case. On average, however, the MPI worked very well. Results from the CAT

pilot indicated problems with scoring the constructed response items in the MATE software. For this reason, constructed response items were eliminated. The final item banks thus included 65 reading items, 105 mathematics items, and 94 science items.

CAT Simulation Study

A second monte-carlo simulation was conducted after the CAT pilot study using the MATE software. This was a rapidly conducted task, given that monte-carlo simulations can be carried out directly, using the item parameters and test specifications from the empirical CAT pilot study. In addition to the item parameters from the final item banks, the θ distributions for reading, mathematics, and science were assumed to follow a normal distribution, each with means and variances as observed in the CAT pilot study. This simulation study also used the empirical item parameter estimates from the final item bank and distributional assumptions from the calibration study (normal distribution with means and variances equal to those in the empirical $\hat{\theta}$ distributions) to simulate new CAT response patterns and to compute estimates of measurement precision and reliability based on these simulated response patterns. The aim of this study was to confirm the results for the reliability estimates obtained from the CAT pilot study, to generate tables of reliability estimates that could later be made available to each of the research projects, and to provide suggestions on the length of the tests, given that these estimates were known.

The CAT algorithm was specified in the same way as in the CAT pilot study, except that the time limit was deactivated and reliability estimates were computed for a given test length. The number of simulees in the three simulations based on item banks and θ distributions designed for the assessment of reading, mathematics, and science competence was N = 1,000. Since the true θs were known in this simulation, the reliability estimates were directly computed as the correlation of $\hat{\theta}$ and θ:

$$\rho(\hat{\theta},\theta)^2_{\text{sim}} = r\left(\hat{\theta},\theta\right)^2 = \left(\frac{\sigma(\hat{\theta},\theta)}{\sigma(\hat{\theta})\sigma(\theta)}\right)^2. \quad [7]$$

The results from this simulation confirmed the previous results on the reliability estimates obtained from the CAT pilot. In each of the three domains, an estimate of $\rho(\hat{\theta},\theta)^2_{\text{sim}} > .70$ was reached after 12 items, on average, were answered. The required testing time would be 20 minutes, 13 minutes, and 10 minutes for reading, mathematics, and science, respectively, according to the time on task computed for each item in the CAT pilot study. A reliability of $\rho(\hat{\theta},\theta)^2_{\text{sim}} > .80$ was reached after 21 items, on average, for reading; 20 items, on average, for mathematics; and 22 items, on average, for science. The required testing time would be 35 minutes, 22 minutes, and 18 minutes for reading, mathematics, and science, respectively.

Thus, most results of the simulation study were the same as those observed in the empirical pilot study. This underlines the robustness of the simulation results and proves that they can be used for predicting the target reliabilities of the tests for different test lengths. The only differences were observed for science. These small observed differences were most likely due to a difference in the standard deviation of the $\hat{\theta}$s between the CAT pilot study (SD = 0.77) and the standard deviation of SD = 0.87 from the calibration study, which was used for data generation in the CAT simulation.

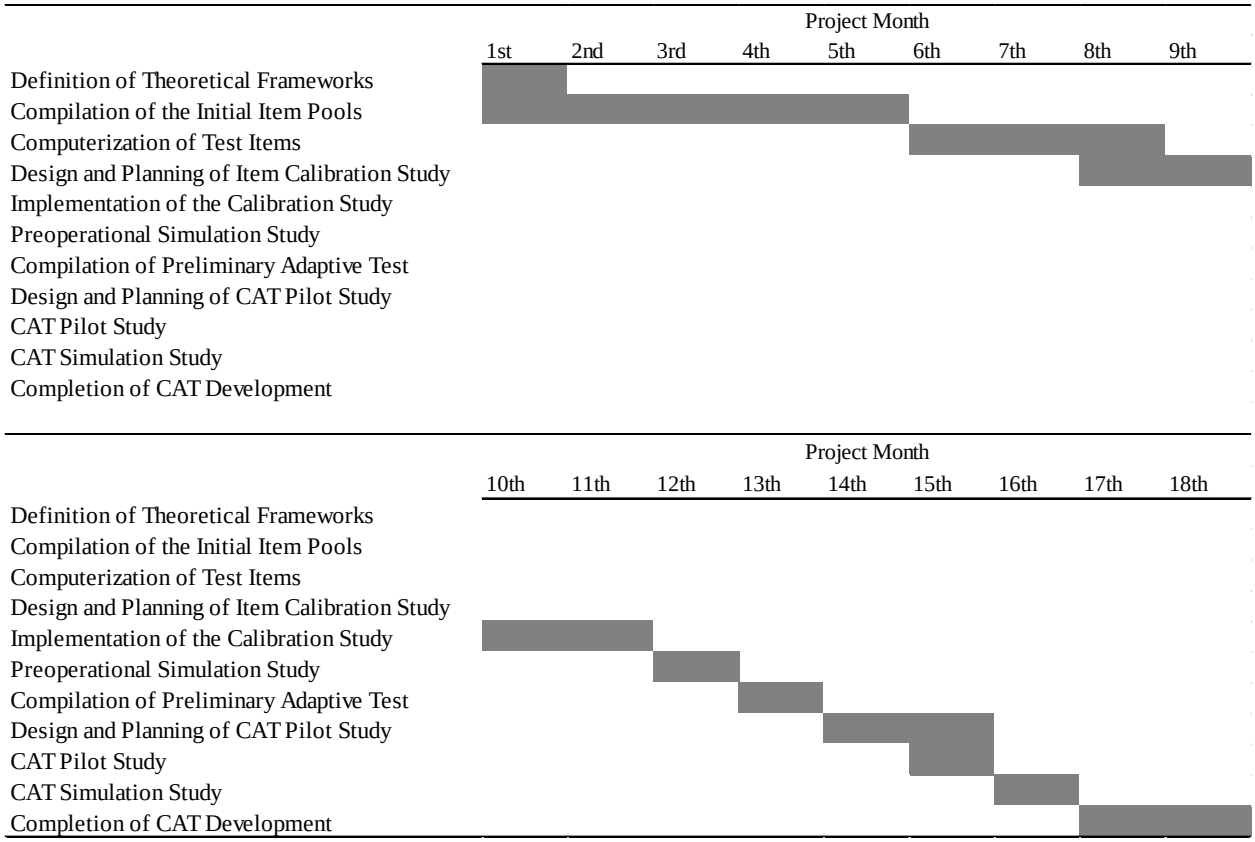
In the project context, the results regarding reliability were used to prepare reliability tables that included the required test length and testing time needed to reach a certain target reliability. The six projects were provided with these tables, enabling them to target the

precision of the measured competencies in reading, mathematics, and science in their studies. Given the successful realization of the CAT pilot and the promising results obtained from the CAT simulation study, each of the three testing instruments was successfully released to be used in the six research projects to measure basic competencies in reading, mathematics, and science.

Timeline of CAT Development

In summary, the complete process of the CAT development required 18 months. The time spent on the different work steps is summarized in the timeline presented in Figure 2. This figure indicates that the compilation of the initial item banks, which was based on the decisions previously made about the theoretical test framework, and the subsequent computerization of the test items were the most labor intensive work steps; these steps were taken early at the beginning of the project. Each of the remaining work steps was realized in no more than two months based on the preparatory work. The design and planning, as well as the implementation of the item calibration and the CAT pilot study, took between one and two months. Both simulation studies conducted were efficiently carried out in approximately one month, using the existing software code in the SAS and MATE software packages. Completion of the CAT development and preparation for test release required an additional two months of work.

Figure 2. Time Required for Different Work Steps in the Project



Conclusions

This study provides two key messages for developers of CATs. First, the development of a CAT that includes item calibration, preoperational simulations, a pilot study, and additional simulations is possible within a short time frame (e.g., 18 months). However, this type of efficient development of an adaptive testing instrument necessitates several favorable conditions. Access to an existing and appropriate theoretical framework is necessary to define the test content, as this step otherwise requires intensive input by content-matter experts. Additionally, a large number of items spanning a wide range of item difficulty needs to be available. For many ability and personality constructs, calibrated item banks are freely available. For the development of CATs measuring student competencies, national and international large-scale educational assessments are a very useful source. Even though appropriate item sets must first be identified and sometimes usage rights have to be requested, the resulting workload is negligible compared to writing new items for the complete item bank. The calibration study should incorporate an elaborated test design, allowing for a balance of items across item positions to balance item position effects.

A further discussion on how item position effects can be accounted for in the item calibration phase can be found in Frey, Bernhardt, and Born (2016). Another requirement for the efficient implementation of CAT is the availability of a software package and an IT infrastructure for running the adaptive test. However, considering the combination of the free *R* software (R Core Team, 2017) and the mirtCAT package (Chalmers, 2016) or the Concerto software (Scalise & Allen, 2015), test developers have access to free, advanced CAT administration software packages in addition to the MATE software.

Second, performing several quality control checks while developing a CAT item bank and piloting the CAT ensures that the testing instrument functions well and is psychometrically sound. This includes using established procedures for item fit analysis and DIF analysis with respect to important person covariates in the item calibration and software testing during a CAT pilot, as well as monte-carlo simulations to examine measurement precision and content balancing. Modern personal computers are easily capable of such simulations. Given a set of item parameters, information regarding the location and the variance of θs, and the predefined constraints on item selection, modern CAT software such as MATE or mirtCAT can implement a simulation module to compute the statistical information of interest from the simulated data.

The CAT development process described here demonstrates the benefits of these simulations. For the present case, they showed that the three CATs worked reasonably well, even with the comparably small item banks. The simulations also underlined that the possible future capacity to add more items should be investigated, especially science items with a high item difficulty (see Figure 1). Results from the simulations further stressed that reliability predictions derived from simulations are only completely accurate if the θ variance can be correctly predicted. If accurate assumptions are not available, the predictions of target reliabilities may be inaccurate.

In general, a rather quick development is an important aim for the successful implementation of a CAT. Its development is typically regarded as a labor-intensive and time-consuming endeavor, indicating a crucial reason for the decision to adopt more traditional paper-and-pencil tests. Thus, the efficient development of CATs should be a major focus of test developers interested in utilizing the benefits of CAT. Quality control processes, however, are necessary to establish a psychometrically sound testing instrument, and they require a certain amount of time. The example presented here illustrates that both objectives—the implementation of an efficient and a psychometrically sound CAT—can be achieved simultaneously.

References

Adams, R. J., Wu, M. L. & Wilson, M. R. (2012). ACER ConQuest (Version 3.0) [Generalised item response modelling software]. Melbourne, Australia: ACER Press.

Bernhardt, R. (2017). *Konstruktion computerisierter adaptiver Tests am Beispiel der Messung schulisch erworbener Kompetenzen* [Measurement of basic competences as an example for the development of computerized adaptive tests] (Dissertation, Friedrich-Schiller-Universität Jena, Germany). Retrieved from https://www.db-thueringen.de/receive/dbt_mods_00031826

Born, S., & Frey, A. (2017). Heuristic constraint management methods in multidimensional adaptive testing. *Educational and Psychological Measurement, 77*(2), 241-262. *CrossRef*

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. *CrossRef*

Cheng, Y., & Chang, H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62*(Pt 2), 369-383. *CrossRef*

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Publications, Inc.

Debeer D., & Janssen R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*(2), 164-185. *CrossRef*

Engelhard, G. (2009). Using item response theory and model data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement, 69*(4), 585-602. *CrossRef*

Frey, A., Bernhardt, R., & Born, S. (2017). Umgang mit Itempositionseffekten bei der Entwicklung computerisierter adaptiver Tests [Handling of item positions effects in the development of computerized adaptive tests]. *Diagnostica*, *63*(3), 167-178. *CrossRef*

Frey, A., Hartig, J., & Rupp, A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, *28*(3), 39-53. *CrossRef*

Frey, A., Heinze, A., Mildner, D., Hochweber, J., & Asseburg, R. (2010). Mathematische Kompetenz von PISA 2003 bis PISA 2009 [Mathematical competence from PISA 2003 to PISA 2009]. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider, & P. Stanat (Hrsg.), *PISA 2009: Bilanz nach einem Jahrzehnt* [Pisa 2009: Balance after a decade] (pp. 153-176). Münster, Germany: Waxmann Verlag GmbH.

Garden, R.A., & Orpwood, G. (1996). Development of the TIMSS achievement tests. In M. O. Martin & D. L. Kelly (Eds.), *Third international mathematics and science study (TIMSS) technical report, Vol. I: Design and development*. Chestnut Hill, MA: Boston College.

Giesbrecht, F. G., & Gumpertz, M. L. (2004). *Planning, construction, and statistical analysis of comparative experiments*. Hoboken, NJ: John Wiley & Sons, Inc.

Helm, C. (2014). COoperative open learning in commercial education: Multilevel analysis of grade 9 students' learning outcomes. *Reflecting Education, 9*(2), *63-84.* Retrieved from http://www.reflectingeducation.net/index.php/reflecting/article/view/129/136

Helm, C. (2015). Determinants of competence development in accounting in upper secondary education. *Empirical Research in Vocational Education and Training*, *7*(10). *CrossRef*

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Kim, S. (2012). A note on reliability coefficients for item response model-based ability estimates. *Psychometrika*, *77*(1), 153–162. *CrossRef*

Konsortium HarmoS Naturwissenschaften+ (2010). *Naturwissenschaften Wissenschaftlicher Kurzbericht und Kompetenzmodell.* Provisorische Fassung (vor Verabschiedung der Standards) [Competence model and suggestions for educational standards for science. Provisional version (prior to adoption of standards)]. Retrieved from https://phzh.ch/MAP_DataStore/158541/publications/HarmoS_2009_Kurzbericht.pdf

Kröhne, U., & Frey, A. (2013). *Multidimensional adaptive testing environment (MATE) manual.* Frankfurt, Germany: German Institute for International Educational Research.

Lane, S., Raymond, M. R., Haladyna, T. M., & Downing, S. M. (2015). Test development process. In S. Lane, M. R. Raymond & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 3-18). London, England: Routledge.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Mullis, I.V.S. & Martin, M.O. (Eds.) (2013). *TIMSS 2015 assessment frameworks.* Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/frameworks.html

Mullis, I.V.S., Martin, M.O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Nagy, G., Nagengast, B., Frey, A., Becker, M., & Rose, N. (2018). A multilevel study of position effects in PISA achievement tests: Student- and school-level predictors in the German tracked school system. *Assessment in Education: Principles, Policy & Practice*. Advance online publication. CrossRef

OECD. (2014). *PISA 2012 results: What students know and can do: Student performance in mathematics, reading and science* (Rev. ed., Vol. I). Paris, France: OECD Publishing.

OECD. (2016). *PISA 2015 Assessment and analytical framework: Science, reading, mathematics and financial literacy*. Paris, France: OECD Publishing.

OECD. (2009). *PISA 2009 assessment framework: Key competencies in reading, mathematics and science*. Paris, France: OECD Publishing.

Preece, D. A. (1996). Youden Squares. In C. J. Colbourn & J. H. Dinitz (Eds.), *The CRC handbook of combinatorik designs* (pp. 511-515). Boca Raton, FL: CRC Press.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* (2nd ed.) Chicago, IL: University of Chicago Press. (Original work published 1960)

SAS Institute Inc. (2011). *Base SAS® 9.3 procedures guide*: *Statistical procedures*. Cary, NC: SAS Institute Inc.

Scalise, K., & Allen, D. D. (2015). Use of open-source software for adaptive measurement: Concerto as an R-based computer adaptive development and delivery platform. *British Journal of Mathematical and Statistical Psychology, 68*(3), 478–496. CrossRef

Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning & Instruction, 13*(2), 141-156. CrossRef

Seeber S., & Lehmann R. (2013). Basic competencies as determinants of success in commercial apprenticeships. In K. Beck & O. Zlatkin-Troitschanskaia (Eds.), *From diagnostics to learning success* (pp. 75–84). Rotterdam, The Netherlands: Sense Publishers.

Spoden, C., Frey, A., Bernhardt, R., Seeber, S., Balkenhol, A., & Ziegler, B. (2015). Differenzielle Domänen- und Itemeffekte zwischen Ausbildungsberufen bei der Erfassung allgemeiner schulischer Kompetenzen von Berufsschülerinnen und Berufsschülern [Differential domain and item effects between professions in measuring basic knowledge

competence of vocational education trainees]. *Zeitschrift für Berufs- und Wirtschaft-pädagogik, 111*, 168-188.

van der Linden, W. J., & Glas, C. A. W. (Eds.) (2003). *Computerized adaptive testing: Theory and practice.* Dordrecht, The Netherlands: Kluwer Academic Publishers

Wilson, M. (2003) On choosing a model for measuring. *Methods of Psychological Research - Online 8*(3), 1-22. Retrieved from https://www.dgps.de/fachgruppen/methoden/mpronline/issue21/mpr122_8.pdf

Wright, B. D. (1980). Afterword. In G. Rasch (Ed.), *Probabilistic models for some intelligence and attainment tests: With foreword and afterword by Benjamin D. Wright*. Chicago, IL: MESA Press.

Wu, M., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ACER ConQuest (Version 2.0) [Generalised item response modelling software]. Camberwell, VIC, Australia: ACER Press.

Ziegler, B., Balkenhol, A., Keimes, C., & Rexing, V. (2012). Diagnostik „funktionaler Lesekompetenz“ [The diagnosis of “functional reading competence”]. bwp@*berufs und wirtschaftspädagogik - online, 22*, 1–19. Retrieved from http://www.bwpat.de/content/ausgabe/22/ziegler-etal/index.html

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Acknowledgments

The preparation of this article was supported in part by Grant 01DB1104 (MaK-adapt) from the German Federal Ministry of Education and Research (BMBF) within the initiative “Innovative skills and competence assessment to support vocational education and training” (ASCOT) and by Grant 16DHL1005 (KAT-HS) from the German Federal Ministry of Education and Research (BMBF) within the initiative “Research on digital higher education.”

Author Information

Christian Spoden is now at the German Institute for Adult Education, Leibniz Centre for Lifelong Learning, Bonn, Heinemannstr. 12-14, 53175 Bonn, Germany. Email: spoden@die-bonn.de. Andreas Frey is now at Goethe University Frankfurt, Germany.