Received 1 September 2015
Revised 9 September 2015
Accepted 10 September 2015

Increasing testing efficiency through the development of an IT-based adaptive testing tool for competency measurement

Janne Kleinhans and Matthias Schumann
University of Goettingen, Goettingen, Germany

Abstract

Purpose – This paper investigates the potential of computerized adaptive testing for CMs to reduce test time. In the context of education and training, competency measurement (CM) is a central challenge in competency management. For complex CMs, a compromise must be addressed between the time available and the quality of the measurements. Increasing the efficiency of existing tests therefore poses a key challenge.

Design/methodology/approach – Results from a market analysis show a lack of integrated adaptive testing tools for CMs. The implementation, integration and evaluation of an appropriate adaptive component are presented for the example of the learning management system ILIAS used for a CM of health workers. The textbook scenario of a linear testing implementation is compared with results from the adaptive testing tool implementation. A simulation component is created to investigate the robustness of the adaptive test regarding answers that are inconsistent with the test person's ability.

Findings – A test time reduction of 40 per cent is achieved by the adaptive test. The test shows high stability for up to 20 per cent of the answers being inconsistent with the test person's ability.

Originality/value – Results indicate the high potential of computerized adaptive testing for CMs. The developed tool could be used for the implementation of combined multimedia computerized adaptive tests. Time savings could be utilized to improve measurement quality. Large-scale empirical studies on the interaction between competency dimensions could be facilitated. No other tool could be identified possessing these capabilities.

Keywords Assessment, E-Learning, Adaptive testing, Competence measurement, Learning management system, Testing efficiency

Paper type Research paper

1. Problem Statement

Competence measurement (CM) is a central challenge in competence management. The more exact competencies can be measured before a training, the more targeted a training can be (Klett, 2010; North *et al.*, 2013; Draganidis and Mentzas, 2006). A formative or summative assessment following trainings and lectures offers the foundation for monitoring learning progress. CMs can vary broadly in their application, ranging from large-scale assessments (e.g. the PISA study) to self-testing.

CMs are characterized by their large scope. There are typically many dimensions to consider that, in turn, depend on numerous characteristics of competency. Erpenbeck and Sauter (2013) differentiate various competency classes, including personal, activity and decision-making, subject and method and social-communicative. According to Heyse and Erpenbeck (2007), these can be divided into facets-like personal

Interactive Technology and Smart Education
Vol. 12 No. 4, 2015
pp. 242-255

1741-5659
DOI 10.1108/ITSE-09-2015-0023

responsibility, decision-making ability and analytical or communication skills. The measurement task grows with every added dimension. If, for example, the mathematical and linguistic abilities of a test subject are to be assessed, it is usually necessary to pose a distinct task for each. Tests cannot, however, be extended to assess an arbitrary number of areas due to the fatigue of test subjects, to the opportunity cost to the examiner (or test subject) or due to testing regulations. A reduction in measurement quality or dimensionality must, therefore, be often accepted in current practice.

The question arises as to how the efficiency of actual CM test procedures can be improved and, in particular, as to how the measurement quality can be raised for a test of fixed duration. Linear testing procedures present an important point for improvement. There are weaknesses in the efficiencies of linear tests of participant groups of heterogeneous abilities, as the tests must remain comprehensive to their groups. This is reflected, for example, by the use of tasks of varying difficulty. At the same time, participants must be presented with the same tasks, thereby being measured at a level that can be either too high or too low. Here, computerized adaptive testing promises significant improvements in efficiency by being able to adapt to the individual levels of participants as a test progresses. There is, however, a lack of tools to facilitate integrated adaptive competence measurements. This is the motivation of this paper. At the example of a CM for health workers, it will be investigated to what extend a computerized adaptive test (CAT) can increase the efficiency of a CM. An appropriate tool will be implemented and evaluated. The following Research Questions follow:

RQ1. How must a tool for computerized adaptive competence measurement be constructed?

RQ2. Can a CAT increase the measurement efficiency of a competence assessment?

Remarks on computerized adaptive testing will follow in Section 2. In Section 3, the requirements of an integrated tool for adaptive competence measurements will be presented. The system choices for the implementation of the CM for health workers will be discussed there. The implementation will be presented in Section 4. The developed tool will be evaluated in Section 5. The discussion of the results comprises the conclusion in Section 6.

2. Computerized adaptive testing

Computerized adaptive testing dates back to the 1970s, when powerful computers became increasingly available (Lord, 1976, 1980; Weiss, 1982). Numerous publications have more recently addressed specific aspects of adaptive testing, like impacts on participants and motivation (Frey *et al.*, 2009; Tonidandel *et al.*, 2002) or content balancing strategies (Zheng *et al.*, 2013), while others provide comprehensive considerations (Van der Linden and Glas, 2000).

The special feature of CATs is that the compilation of their tasks (items), and thus the level of the exam, is first established during the test and is dependent on the ongoing performance of the tested person. A more difficult task will typically follow a correct answer, and vice versa. For the purposes of this, each item has a numeric value representing its degree of difficulty. These values are usually determined in a calibration phase, during which the tasks are given to a comprehensive group of test subjects whose performance is factored into the statistical models presented below. CATs thus provide an individual testing experience to each participant and thereby

increase testing efficiency as compared to with classical linear methods, as items deemed too easy or difficult can be excluded and each item becomes diagnostically useful.

CATs aim for the maximum possible performance of the test person. Their goal is to adapt to problems for which the examinee has a 50 per cent success rate. Unlike for a linear test, performance cannot be measured by the total number of correctly answered questions, due to the varying difficulty of the questions. For this reason, CAT is bound by Item Response Theory (IRT), which allows for an assessment of performance based on the answered tasks rather than on the test itself (Baker and Kim, 2004). The numerical ability parameter (θ) reflects the competency level of the participant and replaces the relative number of correct answered questions as the test result. The standard error of the test result is calculated in real time and provides a measure of the accuracy of the assessment.

CATs must be separated into branched and tailored types. While for the former a branch of questions is a predetermined function of the participant's answers, for the latter the participant's testing level is recalculated with each answer and the subsequent questions are chosen accordingly from all available tasks in the item pool. This offers greater efficiency (Kubinger, 2009).

3. Implementation criteria and software identification

The requirements for an IT-based integrated CM tool will be defined in this section. General requirements and criteria from the applied sample CM for health workers are mentioned at first. The system selection for the CM implementation follows.

Both open and closed questions are of interest for CMs. Closed questions like multiple-choice ones facilitate an economical assessment of knowledge due to their simple structure. More complex types can address functional capacity (e.g. requiring the examinee to select the proper region of a figure). Open questions, like free text entries, can assess complex skills (like communication or problem-solving skills, e.g. through the structuring of answers). The possibility for both open and closed questions is, therefore, a necessary criterion for task creation. Multimedia-based elements can add significant value to a CM by increasing the action of the relevant task. They can present complex stimuli (Brunken *et al.*, 2003) that support the situational and contextual integration of the tasks and likewise promote the transferability of the test results to real-world situations (Jurecka and Hartig, 2007; Mayer, 2005). The availability of multimedia elements is, therefore, another required criterion. As tailored testing offers the greatest potential for increasing testing efficiency, the availability of an adaptive component that supports tailored testing is also required.

After these general requirements, criteria for the use as testing software for health workers will be derived. The vocational training of health workers in Germany follows a dual-study system. The training takes place in both a medical practice and a vocational school and requires the completion of an intermediate and final examination. The tool should work for both a large-scale summative assessment in the vocational school and an assignment in the medical practice, e.g. a formative self-test. Consequently, the test tool should be highly scalable and applicable under heterogeneous conditions. These requirements could also be applied as a typical scenario for many other professions. At the same time, the evaluation must be able to provide feedback to the tested persons. The criterion of an intuitive user interface is

necessary to minimize barriers for test takers not having high levels of computer literacy. As the final examination of the health workers is a summative assessment, it is essential for the later evaluation that the results are classifiable. A functionality for the archiving of test results is necessary. The encryption of data transfers and secure data storage address requirements for the testing security as set by the German federal states. Meeting these requirements ensures that the tool could be used for examinations or training purposes in future.

The test should be split into adaptive and multimedia-based parts, such that the efficiency of the CAT could be optimized regarding its layout. Multiple-choice questions without multimedia elements ensure that the test remains as simple as possible. For the multimedia-based part, graphics, videos and free text tasks are required to increase the activeness. The multimedia-based test will not be discussed here, as this paper focuses on the increase to the test efficiency through CAT.

To identify a suitable software solution, a market analysis was conducted. Following Webster and Watson (2002), a literature search was carried out using the keywords e-assessment, computer-based assessment, computerized assessment and computer-based testing, in both English and German. Through the search engines Google and Bing, 136 potential software solutions were subsequently identified.

The suitability of the various potential solutions was examined in three steps with increasingly particular criteria. This division helped to avoid potentially viable solutions from being prematurely excluded. This proved beneficial, as no tool was identified that met all criteria, and a compromise was required for the further implementation. In total, 65 solutions were excluded in the first filtering step due to a lack of a graphical user interface, an inability to assign test results to subjects, to internally assess answers as right or wrong, or an inability to store test results long-term. In filtering Step two, 63 solutions were excluded, as they were restricted to specialized use cases (e.g. mathematical tests) and did not allow for usage in the case of health workers, allowed no possibility for self-hosting or encrypted file transfer or offered incomplete support of multimedia elements. Two of the remaining eight solutions did not support all necessary question types and all of them lacked an adaptive component for tailored testing.

Because no software could be found that satisfied all criteria, focus was turned to finding the most expandable solution. A detailed analysis showed that no specialized adaptive solutions offered a suitable basis for expansion, as all failed to satisfy several criteria, like the possibility to integrate multimedia elements, shortcomings in testing security and test management. As six solutions met all criteria except for the adaptive testing functionality, they were examined more closely. The analysis was performed using the categories of issue management, test management, security, interoperability, usability and reliability.

On this basis, the test component of the learning management system ILIAS (ILIAS, 2015; Kunkel, 2011) was selected for further development. As an open-source software in a standard programming language, PHP, ILIAS offered complete freedom for customization and the long-term availability. Being Web-based and platform-independent, the solution satisfied the requirement for usability under heterogeneous conditions. The client server structure allowed for the central storage of all test data. In the event of a system crash, tests can be resumed from the appropriate place. There was also high scalability with support for extensive user management and numerous

simultaneous users. ILIAS offers 11 question types. The analysis is largely omitted and allows for summative and formative efforts.

4. Implementation

In Section 4.1, the concrete requirements for the implementation are described based on the functional criteria from Section 3. The CM realization is presented in paragraph 4.2.

4.1 Requirements

The CAT should diverge as little as possible from other ILIAS functionalities to avoid barriers to the usability. For maximum compatibility, the CAT should work with the same system requirements than standard (ILIAS, 2015; Kunkel, 2011). Solely, a limited increase in computing power will be necessary due to the CAT algorithm. However, it should be limited as the test has to be usable under heterogeneous conditions in the vocational schools or medical practice.

The CAT should be based on IRT. By using IRT, it is possible to connect the test results with one or more latent variables on an empirical foundation (Baker and Kim, 2004; Embretson and Reise, 2000). In the current case of performance testing, one latent variable is considered to be the participants' ability to solve the test items (θ), which is consequently represented as a latent trait. The modeling of the corresponding representation, which allows the statistical calculation of the latent trait, should be done using the Rasch model. It assumes that a person's ability to solve a test item is based on the persons' latent trait – represented by the estimation of a weighted likelihood estimate – and the difficulty of the item – represented by a response model parameter estimate. Both parameters are estimated based on solution probabilities and are interdependent in iterations (Bond and Fox, 2007; Fischer and Molenaar, 1995). According to the Rasch model, whether a person solves a problem depends only on his or her ability and on the difficulty of the task, which are both measured on the same scale. This is a strong assumption, with the advantage of facilitating the modeling of the adaptive test solely on these parameters.

An extensive calibration phase (cp. 5.1) was conducted to determine the difficulty of the tasks and decide whether a one- or multi-dimensional competency model should be used. A one-dimensional model was selected, as multi-dimensional models showed no significant increase in precision. However, the possibility to expand the tool later for multi-dimensional CMs should exist. There are various one-dimensional estimation methods that implement IRT (Van der Linden and Glas, 2000). Because the tasks were predominantly developed from scratch and there existed no prior experience in the calibration of the CAT, the simplicity and robustness of the estimation method are a priority, given the complexity of the procedure and the precision of the measurements. The expansion for adaptive testing was divided into three parts, namely:

(1) the creation of tasks;

(2) the processing sequence during the test procedure; and

(3) the storing of the test data and the evaluation of the test results.

For the creation of the adaptive tasks (1) above, the ILIAS task template must be extended. The difficulty of each task needs to be added as numerical value to each question type. A second numerical parameter should be created to assign different competency dimensions to the tasks for a potential later multi-dimensional CM.

Furthermore, the possibility should exist to deploy a task as adaptive or non-adaptive in different tests. Therefore an additional Boolean parameter should be added.

The processing sequence during test procedure (2) above in case of a CAT must be constructed as the test progresses. In contrast, ILIAS uses for sequencing tasks an initial test sequence, which is not changed during the test. Even though ILIAS offers an option for randomized item selection, only the initial sequence is generated randomly in this case. Therefore, a constant reordering of this test sequence during the test has to be implemented. There are three determining factors for the adaptive testing procedure. These include:

(a) item selection during the test;

(b) item selection at the start of the test; and

(c) the conditions for the completion of the test (Van der Linden and Glas, 2000; Mills and Stocking, 1996).

All factors depend directly on the used estimation method. The expected a posteriori (EAP) approximation of Bock and Mislevy (1982) is chosen, combining decent precision with high robustness and low computing requirements. EAP is a Bayesian method. For item selection during the test after each completed task, the ability parameter of the participant (θ) is estimated and subsequently the most informative unused item for the current θ (= item, whose difficulty is next to (θ)) is selected. EAP offers a small bias and standard error compared to other Bayesian methods (Wang, 1997). The EAP estimates are calculated noniteratively. Corresponding calculations using values of the IRT function can be performed before the beginning of the test and stored (Bock and Mislevy, 1982), which reduces hardware demands. Another advantage of EAP is that it could also be used for the item selection at the start of the test, as the approximation is stable over the entire test length (Bock and Mislevy, 1982). An average skill level ($\theta = 0$) should initially be assumed. The standard termination condition in ILIAS for linear tests is met, when the last item in the test sequence is passed. For the adaptive test, two termination conditions should be established, the first being the static completion of a defined number of questions and the second being the dynamic achievement of a certain level of precision with regard to the estimate of the standard deviation.

For the storing of the test data and the evaluation of the test (3) above, the data storage has to ensure that the data remain accessible over the long term. For the evaluation, standard functionality needs to be expanded to output the task, the updated measured competency level (θ) and standard error following each question. In addition, an aggregate index of all questions is required for the evaluation of all test persons and tasks.

A simulation component should be constructed to allow the creation of data sets for predetermined θ values and to analyze the performance and robustness of the test. For a list of given person's abilities (θ), the test procedure should be performed automatically. If a person's ability (θ) is higher than or equal to the difficulty level of the question, the question should be answered correctly by the system, if it is wrong, it should be answered incorrectly. Furthermore, there should be a randomization component. This should – based on predetermined probabilities – reproduce the chance in the simulation that a strong participant answers a question wrong, even if his or her

θ is higher than the difficulty value of the question, or that a weak participant guesses a question right, even if his or her θ is lower than the difficulty level of the question.

4.2 CM realization

The ILIAS task template was completed for each questions type to establish a level of difficulty, a subject area and for the identification of adaptive questions. The user can, consequently, create or edit adaptive tasks through the standard interface without programming skills.

The CAT algorithm is selected over standard algorithms through a new option in the interface. When a question pool is selected by the user for the creation of an adaptive test, all adaptive questions within it are used to create an initial test sequence. This ensures that all items are available for the test procedure. During the execution of the test, this sequence is continuously refined: after each completed task, the EAP estimation is applied to select the next item. The selected item is promoted to the position which is next in the test sequence. The standard error is calculated at the same time with θ. At the beginning of the test a value of $\theta = 0$ is presumed. A question number limit and a criterion for the standard error in the algorithm are possible termination conditions to be used independently or in tandem. The inputs for the termination conditions were integrated into the standard interface. When termination conditions are met, the test sequence is shortened to this point and the standard termination condition of ILIAS is used to finish the test.

The test results (θ and standard error) are continuously stored in the central ILIAS database so that the CAT can be resumed after a system crash without any loss of progress. The corresponding data and the data from all newly created input fields (e.g. difficulty of questions, termination conditions) are archived in a way consistent with the other data. For the evaluation, two new spreadsheets were created. One table shows the individual testing trajectory of the participant, including his or her skill level and the standard error. The second gives an overview of all questions posed to at least one user and how they were answered.

All possibilities for the multimedia design were retained for the CAT. This is a clear benefit in comparison to numerous special CAT software solutions. The encapsulated design of the testing tool facilitates the integration of additional functionalities, e.g. multi-dimensional estimation procedures.

A simulation component was constructed and integrated as a new menu item in the test results template. The user can upload a .csv list with all θ values he or she wishes to simulate. The test can then be performed according to these values automatically. If the standard termination conditions of the test should not be used for the simulation, they can be changed exclusively for this purpose. Furthermore, two fields were created to simulate the chance of failure even when the individual θ is higher than the difficulty of the task and vice versa. Both are entered as per cent values. For the evaluation of the simulated data, the same spreadsheets as for an adaptive test with normal test takers can be used.

5. Evaluation

This section presents the evaluation of the testing instrument. The tool was calibrated and evaluated in a large-scale scenario with real scholars in the vocational schools. A description of the general experience concerning the application is given first. A

comparison of concrete measurements with those of a typical linear test is found afterwards.

5.1 Quality assurance and experience with the implementation

The newly developed tasks were tested in two ways. In the first calibration stage, to establish the degree of difficulty of each task, they were presented linearly to approximately 1.200 health worker scholars in the vocational schools. The graduating class was chosen for maximum comparability of the results for the final examination. Furthermore, they were examined in expert workshops. The usability of the test proved to be very good. Because the surface of the CAT did not differ from a linear standard ILIAS test, a bias due to a different interface could be excluded. The items were coded dichotomous. Partially correct solutions were evaluated as false to increase their difficulty. After final revisions, scaling and testing of Rasch conformity, 88 items covering a skill range (θ) of -2.091 to 2.678 remained.

In a second stage, these items were used in main survey for a span of 15-30 questions. As no detailed experiences regarding the expected precision were available, a fixed number of questions was chosen in favor of a variable test length. A total of 1,183 data sets were collected through adaptive testing. These data sets were measured using participants that differed from the first stage.

5.2 Comparison of the linear and adaptive testing procedures

To check whether the adaptive design of the tests could increase the measurement efficiency, a textbook example of a linear test procedure was created as a basis for comparison. Five people with a skill level spanning from $+2$ (high) to -2 (low) were taken with a test length of 15 questions equally divided into five difficulty levels. While the specification of skill levels in CATs depends on the statistical model, it is specified here manually to provide integer number levels of competency. The scale is interpolated to the value range of the CAT for comparison. Even if 93.5 per cent of the participants had a final skill level between -1.4 and 1.3 in the CAT, the range from -2 to 2 was selected to examine the full range of the CAT (best and worst results were 2.21 and 1.97, respectively). The textbook example is shown in Figure 1.

The number of individual test questions is presented in the first line, the difficulty level of the tasks in the second and the results for the individual test persons in the five following lines. The tests were completed from left to right. It is assumed that the difficulty of the tasks is the only influencing factor, such that a respondent properly

task number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	θ test taker = θ task	no contribution to measurement
θ task	-2	-2	-2	-1	-1	-1	0	0	0	1	1	1	2	2	2		
test taker θ = +2	< θ	< θ	< θ	< θ	< θ	< θ	< θ	< θ	< θ	< θ	< θ	< θ	= θ	= θ	= θ	20,0%	80,0%
test taker θ = +1	< θ	< θ	< θ	< θ	< θ	< θ	< θ	< θ	< θ	= θ	= θ	= θ	> θ	> θ	> θ	20,0%	60,0%
test taker θ = 0	< θ	< θ	< θ	< θ	< θ	< θ	= θ	= θ	= θ	> θ	> θ	> θ	> θ	> θ	> θ	20,0%	60,0%
test taker θ = -1	< θ	< θ	< θ	= θ	= θ	= θ	> θ	> θ	> θ	> θ	> θ	> θ	> θ	> θ	> θ	20,0%	60,0%
test taker θ = -2	= θ	= θ	= θ	> θ	> θ	> θ	> θ	> θ	> θ	> θ	> θ	> θ	> θ	> θ	> θ	20,0%	60,0%

Figure 1.
Linear testing order

completes all tasks of the same or lower skill level (test person $\theta >=$ task θ). No distinctions are made according to the content areas. A green cell marking signifies that a task has the same skill level as the test person, while a yellow one signifies that it was taken as a bound. The last two columns present the portion of tasks at the skill level of the test person and the measurements of irrelevant tasks (red marking).

For all five sample test persons, the individual skill level was only matched for 20 per cent of the tasks. The strongest test person ($\theta = 2$) initially needed 12 items below his or her skill level before receiving items 13-15 at this level. For the weakest test person ($\theta = -2$), only the first three tasks were on the appropriate level and questions 4-15 were too difficult. The other cases performed equivalently, with only the position of the task fitting the test person's skill level changing. Except for the strongest test person, at least one further task was necessary to delineate the upper limit of the skill level. It could be generously argued that the entire next level of difficulty is relevant to distinguishing the skill limit (light marking), such that 20-40 per cent of the questions add value to the measurement. In reverse, for the present example, 60-80 per cent of the questions (no marking) have no direct contribution to the measurement, as they are too easy or difficult. This rate could be reduced with the difficulty level value. Apart from the loss of precision, however, the basic problem would persist.

To test whether improvements can be achieved through the CAT for this idealized scenario, the scenario was adopted and supported by measured values from the CAT application. Based on the stepped skill levels +2, +1, 0, −1 and −2, one of the 1,183 measured data sets was chosen for each test person based on which most nearly fit his or her measured θ as evaluated after 15 questions. Test persons with θ values of +1.95, +1.00, 0.00, −0.99 and −1.97 were thereby identified. Figure 2 illustrates the testing

task number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	θ test taker = estimated θ	no contribution to measurement
test taker θ = + 1,95 (+1,5 to +2,5)	0,41	0,76	1,07	1,34	1,59	1,82	2,04	2,20	2,05	2,16	1,91	2,00	1,80	1,88	1,95	73,3%	0,0%
test taker θ = + 1,00 (+0,5 to +1,5)	0,41	0,06	-0,24	0,02	0,26	0,47	0,67	0,85	1,03	0,86	1,01	0,87	1,01	0,88	1,00	60,0%	0,0%
test taker θ = 0,00 (-0,5 to +0,5)	-0,41	-0,06	0,24	-0,03	-0,26	-0,05	-0,24	-0,42	-0,25	-0,40	-0,26	-0,13	-0,01	0,11	0,00	100,0%	0,0%
test taker θ = - 0,99 (-1,5 to −0,5)	-0,41	-0,06	0,24	-0,03	-0,26	-0,47	-0,67	-0,85	-0,68	-0,83	-0,98	-1,12	-0,99	-1,11	-0,99	93,3%	0,0%
test taker θ = − 1,97 (−2,5 to −1,5)	-0,41	-0,76	-1,07	-1,34	-1,59	-1,82	-1,60	-1,78	-1,63	-1,77	-1,89	-2,00	-2,09	-1,90	-1,97	73,3%	0,0%

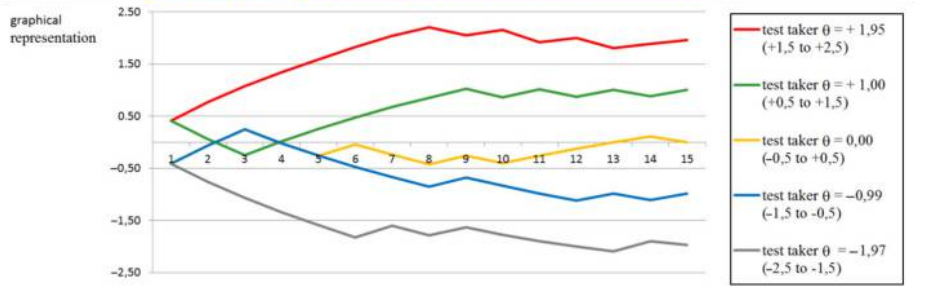

Figure 2.
Adaptive testing order

procedure for these test persons, showing the measured skill values following each question.

The presentation follows the likeness of the previous figure. A graphical representation of the test process is added for better understanding. Each participant answered 15 questions, numbered left to right. Due to the adaptive nature of the test, the questions differed between participants. As there can be no universal statement of the difficulty of the tasks, the second line is omitted. It must be noted that – even if both are measured on the same scale – there is no one-to-one correspondence between the measured test result and the presented item. Especially for very high or low skill levels ($\theta = \pm 2$) with a lower density of items differences can occur, when the estimation process needs to compensate for items that do not exactly fit. The estimated skill level, therefore, offers more precise information of the current state of the estimation than the difficulty of the task and is used for further descriptions. In the linear textbook example, it was taken into account whether the difficulty of the task corresponded to the participant's θ value. As in CAT, where the estimation result following the previous task is the basis for the choice of the current task, it would be plausible to use the estimation result of the previous task. However, this would neglect the increase in precision that was already achieved through the current task. Therefore, for each question number in Figure 2, the absolute measurement result (i.e. the test person's θ to two decimal places) following each question is presented to determine whether a task addressed the skill level of a participant. An interval of $\theta \pm 0.5$ is taken as a basis for comparison. This would, in the worst case, correspond to the same accuracy limit of the linear textbook scenario: for the textbook example, the difficulty level must exceed the participant's skill level by 1 (e.g. to $\theta = 2$ for a participant with $\theta = 1$), whereas for the CAT an interval extending both above and below, the participant's skill level is necessary (e.g. the interval $\theta = 0.5$ to $\theta = 1.5$ for a participant with $\theta = 1$). Tasks for which the measurement results addressed the final skill levels of the test persons are characterized by green cell markers. Yellow highlights are used to distinguish contributing tasks to measurement.

It is apparent that between 60 and 100 per cent of the questions addressed the relevant skill level. The share of too easy or difficult questions is reduced to 0-40 per cent. The portion of questions not affecting the measurement is reduced to 0 per cent, as each question served part of the estimation process and the determination of θ. As expected, the border values fall weakly for high and low θ, as the algorithm takes longer to settle on the appropriate skill level. It becomes clear, however, that a skill level within the appropriate interval is reached after seven consecutive questions and that a nearly constant θ is reached after nine questions. This is delineated by the vertical dashed line in Figure 2. With Question 9, each test person was given at least three tasks directly affecting his or her own skill level as for the linear test. The black vertical line marks this. For the presented example, at least 60 per cent of the questions (after Question 9 instead of 15) would have already been reached at this point. The desired increase in the test efficiency due to the CAT algorithm could be confirmed by the example case. The mean value of θ was calculated over all 1,183 collected data sets after Questions 9 and 15. Both values were very similar (0.261 versus 0.296), highlighting the potential increase in testing efficiency. At the same time, the positive value of θ shows that the CAT was slightly easier than expected.

Additionally, the example of the two intermediate test persons (cp. Figure 2) illustrates that the algorithm is from the start able to quickly change direction. The test person with final $\theta = -1$ guessed the Questions 2 and 3 right, even though his or her final θ was lower than the difficulty of the questions. In contrast, the test person with $\theta = 1$ failed to answer the Questions 2 and 3 correctly despite having a final θ that was higher than the difficulty of the questions. By Question 7, the test however already reaches in both cases the final ability intervals of the two test persons.

To further analyze the influence of answers not corresponding to θ (= ANCθ), a simulation study was performed. The same values from the example above (θ values and corresponding intervals with a test length of 15 questions) were chosen for comparison. The rate of ANCθ was increased in two steps from 0 to 6.67 per cent (corresponding to one of the 15 questions) and 13.33 per cent (corresponding two questions). The ANCθ value is not fixed and varied randomly to assess its general influence. In total, 100 data sets were simulated for each θ value and ANCθ rate. Table I illustrates the results.

In each row, the mean values for all 100 data sets corresponding to the same θ are shown. Columns one and two present the preselected θ value and ANCθ rate, respectively. In column three, the task position for which the final difficulty stabilized is shown.

The simulation data underline that the test takers with $\theta = \pm 1$ from Figure 2 suffered some unexpected answers. In case of corresponding answers to θ for questions two and three, the test would have been able to reach the final difficulty interval already after task two (instead of seven). For the same reason, in contrast to the measured data of $\theta = \pm 1$ presented in Figure 2, the simulation data show that it takes longer to stabilize on the extreme outer values of $\theta = \pm 2$ than for the values of $\theta = \pm 1$, which conforms to expectation that the CAT needs longer to branch to extreme values.

It can be observed that a higher ANCθ rate has a higher influence for the outer skill levels. For a rate of 13.33 per cent, the test length for participants with a θ of 0 only increased by 0.72 questions. In contrast, it is expanded on average by 1.26 questions for $\theta = \pm 1$ and by 3.71 questions for $\theta = \pm 2$.

For higher ANCθ rates, further simulations showed that the test starts to lose stability. For a rate of 20 per cent (corresponding to three of 15 questions) and $\theta = \pm 2$, it can be observed that 7 per cent of the simulated data sets fail to stabilize even if the test length is increased to 30 questions, whereas the test length for other data sets was expanded by 8.5 (to 13.5) questions. For a rate of 26.67 per cent (corresponding 4 of the 15 questions), the test still stabilizes for a for $\theta = \pm 1$ but completely fails for $\theta = \pm 2$. Detailed analysis yields two explanations. As the branch of the algorithm in general

Table I. Influence of unexpected answers

θ	ANCθ rate (= 1/2/3 of 15 questions not answered corresponding to θ)	position of stabilization of test (= final difficulty interval reached and not left again with question)
−2	0%/6.67%/13.33%	5.00/6.67/8.13
−1	0%/6.67%/13.33%	2.00/2.69/3.31
0	0%/6.67%/13.33%	1.00/1.28/1.72
1	0%/6.67%/13.33%	2.00/2.61/3.21
2	0%/6.67%/13.33%	5.00/6.21/9.29

takes longer to reach extreme values, it is also more sensitive to ANCθ. Second, item density is very important. In the current case, 90 per cent of the items factored the range from -1.4 to 1.3. Therefore, if tasks outside this range could not be transferred into expedient information according to ANCθ, there were insufficient items available to recover from the false information. It should, however, be noted that a ANCθ rate of 20 per cent still represented a high level of disturbance. Even in linear test settings the test result would suffer if every fifth questions was answered incongruent to the skill levels.

With regard to the real measured data sets, the question follows as to how often items were used in the test and whether there were any ceiling effects. The number and position of usage is, therefore, analyzed for all items: All test persons are given the same item with intermediate difficulty at the start of the test. This conforms to expectations, as all test persons have the same θ of 0 at the beginning of the test. Dependent on the correct or incorrect answering of the first question, an easier or more difficult subsequent item is selected; accordingly the algorithm branches to 4 items, then 8 and 16. By the ninth item all questions, spanning the entire difficulty spectrum, not posed up to this point, belong to parallel paths. The frequency of the use of items of intermediate difficulty decreases toward the limits of high and low difficulty. However, the more difficult items are more frequently used, confirming the test was easier than expected. A weak ceiling effect is present for strong participants, as certain very difficult items have an increased usage rate in some positions. This effect is, however, limited, as the absolute item usage still decreases recognizable to the upper limit.

6. Conclusion

The central point of this contribution was to examine whether CAT could improve testing efficiency in the example of a medical competency measurement, as well as to present the relevant implementation and functional ability of an CAT component in ILIAS.

The results show that a fully valued CAT was successfully developed for the learning management system ILIAS. The test reacts adaptively to user inputs and selects the necessary subsequent tasks during runtime. The requirements in Section 3 and the implementation in Section 4 provide a detailed answer to research Question 1, how to construct a computerized tool for adaptive competence assessment. Research Question 2, focusing on the increase in measurement efficiency for a competence assessment, was addressed in Section 5. A clear increase in the measurement efficiency could be achieved for the presented implementation case of a CM for health workers. As compared to a traditional testing format, the test time was reduced by 40 per cent. The test persons were already posed three questions addressing their competency level after 9 questions in total, as compared to after 15 total questions for a linear test procedure.

In addition to the 1,183 measured data sets, a simulation was used to investigate the robustness of the test to answers inconsistent with the test person's ability. It was observed that especially for very high- and low-skill levels, such answers necessitated a longer test length for accurate results. For low skill levels, this can be particularly important if the influence of randomly guessing questions correctly must be compensated for. For the developed test, the analysis showed an adequate item density and robustness to such answers.

In summary, an integrated tool was created for the competency measurement, with which a multifaceted adaptive competency measurement can be created from

comprehensive types of questions and multimedia elements. No other tool could be identified possessing these capabilities, including extensive reporting options, graphical interface and high test security. This facilitates the potential for future implementations of combined multimedia-CATs. The tool supports large-scale testing and summative, diagnostic or formative usage. The time savings realized through the implementation of the CAT can be utilized as part of an integrated competency measurement with further testing. In practice measurement quality or dimensionality of the test can be improved for the same participants or more participants could be tested within the same time. The mapping of complex action situations in multimedia tests could replace personnel-intensive oral examinations. This benefit is not limited to CMs in the medical field.

As a contribution to knowledge, along with the time savings, the testing tool could facilitate more detailed and larger-scale competency measurements. It could facilitate large-scale empirical studies on the interaction between competency dimensions that are currently not feasible because of the associated expenses.

This study faces some limitations that should be addressed in further research. The adaptive component has not been used for multimedia adaptive testing and the corresponding potential was not evaluated. An alternative adaptive algorithm has also not been implemented to compare possible effects on the testing efficiency. The simulated answers not conforming to θ were also not related to item position and only the general influence was investigated. They are all points for further work.

References

Baker, F.B. and Kim, S.H. (2004), *Item Response Theory: Parameter Estimation Techniques*, Marcel Dekker, New York, NY.

Bock, R. and Mislevy, J. (1982), "Adaptive EAP estimation of ability in a microcomputer environment", *Applied Psychological Measurement*, Vol. 6 No. 4, pp. 431-444.

Bond, T.G. and Fox, C.M. (2007), *Applying the Rasch Model*, Routledge, Mahwah.

Brunken, R., Plass, J.L. and Leutner, D. (2003), "Direct measurement of cognitive load in multimedia learning", *Educational Psychologist*, Vol. 38 No. 1, pp. 53-61.

Draganidis, F. and Mentzas, G. (2006), "Competency based management: a review of systems and approaches", *Information Management & Computer Security*, Vol. 14 No. 1, pp. 51-64.

Embretson, S.E. and Reise, S.P. (2000), *Item Response Theory for Psychologists*, Lawrence Erlbaum, Mahwah.

Erpenbeck, J. and Sauter, W. (2013), *So Werden Wir Lernen*, Springer Gabler, Berlin Heidelberg.

Fischer, G.H. and Molenaar, I.W. (Eds) (1995), *Rasch Models: Foundations, Recent Developments, and Applications*, Springer, New York, NY.

Frey, A., Hartig, J. and Moosbrugger, H. (2009), "Effekte des adaptiven testens auf die motivation zur testbearbeitung am beispiel des frankfurter adaptiven konzentrationsleitungs-tests", *Diagnostica*, Vol. 55 No. 1, pp. 20-28.

Heyse, V. and Erpenbeck, J. (2007), *KompetenzManagement*, Waxmann, Münster.

ILIAS (2015), "Development", available at: www.ilias.de/docu/ (accessed 24 June 2015).

Jurecka, A. and Hartig, J. (2007), "Computer- und netzwerkbasiertes assessment", *BMBF Forschung (Hrsg.): Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik*, Bonn, pp. 37-48.

Klett, F. (2010), "The interrelationship between quality and competency management – the foundation for innovative training technologies", *Information Technology Based Higher Education and Training (ITHET), 9th International Conference*, *Cappadocia*, pp. 174-178.

Kubinger, K.D. (2009), *Psychologische Diagnostik*, Hogrefe, Göttingen.

Kunkel, M. (2011), *Das offizielle ILIAS 4-Praxisbuch*, Addison-Wesley, München.

Lord, F.M. (1976), "Some likelihood functions found in tailored testing", in Clark, C.L. (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing, Washington, DC*, pp. 79-81.

Lord, F.M. (1980), *Applications of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum, Mahwah.

Mayer, R.E. (2005), *The Cambridge Handbook of Multimedia Learning*, Cambridge University Press, Cambridge.

Mills, C.N. and Stocking, M.L. (1996), "Practical issues in large-scale computerized adaptive testing", *Applied Measurement in Education*, Vol. 9 No. 4, pp. 287-304.

North, K., Reinhardt, K. and Sieber-Suter, B. (2013), *Kompetenzmanagement in der Praxis*, Springer Gabler, Wiesbaden.

Tonidandel, S., Quiñones, M.A. and Adams, A.A. (2002), "Computer-adaptive testing: the impact of test characteristics on perceived performance and test takers' reactions", *Journal of Applied Psychology*, Vol. 87 No. 2, pp. 320-332.

Van der Linden, W.J. and Glas, C.A.W. (2000), *Computerized Adaptive Testing*, Kluwer, Dordrecht.

Wang, T. (1997), "Essentially unbiased EAP estimates in computerized adaptive testing", *Annual Meeting of the American Educational Research Association*, Chicago.

Webster, J. and Watson, R. (2002), "Analyzing the past to prepare for the future", *MISQ*, Vol. 26 No. 2, pp. 13-23.

Weiss, D.J. (1982), "Improving measurement quality and efficiency with adaptive testing", *Applied Psychological Measurement*, Vol. 6 No. 4, pp. 473-492.

Zheng, Y., Chang, C.H. and Chang, H.H. (2013), "Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement", *Quality of Life Research*, Vol. 22 No. 3, pp. 491-499.

Corresponding author
Janne Kleinhans can be contacted at: JKleinh@uni-goettingen.de