FIRST RESULTS OF COMPUTERIZED ADAPTIVE TESTING FOR AN ONLINE PHYSICS TEST

U. C. Müller [1]
Hamburg University of Technology (TUHH), Arbeitsstelle MINTFIT Hamburg (AMH)
Hamburg, Germany
ORCID: 0000-0001-8353-200X

T. Huelmann
University Medical Center Hamburg-Eppendorf (UKE Hamburg)
Hamburg, Germany
ORCID: 0000-0003-0697-8969

M. Haustermann
Universität Hamburg (UHH)
Hamburg, Germany
ORCID: 0000-0001-9437-4086

F. Hamann
University of Technology Hamburg (TUHH)
Hamburg, Germany
ORCID: 0000-0001-7166-1437

E. Bender
Hamburg University of Applied Sciences (HAW)
Hamburg, Germany
ORCID: 0000-0002-4912-3955

D. Sitzmann
Hamburg University of Technology (TUHH), Arbeitsstelle MINTFIT Hamburg (AMH)
Hamburg, Germany
ORCID: 0000-0003-2295-3083

Conference Key Areas: *Physics and Engineering Education, Assessment*
Keywords: *computerized adaptive testing, item response theory, online test, physics, prospective students*

[1] *Corresponding Author*
U. C. Müller
ute.carina.mueller@mintfit.hamburg

doi:10.5821/conference-9788412322262.1273

ABSTRACT

Tests are an essential tool to assess students' ability. In online education these tests are mostly of static nature with the same items for each student. In contrast computerized adaptive testing concepts take into account the information about the test user automatically collected in an online test. The aim is a comparably precise test result with fewer test items (questions). An implementation of such a computerized adaptive test (CAT) is presented here. The adaptation process is based on the precise knowledge of the item parameters, e.g. difficulty, in the item pool. An estimation of the knowledge level of the test user has to be performed in real time after each answer. With this information the next item can be selected accordingly. This leads to a highly individualized test for each test user. For all items the parameters were determined with methods of the item response theory (IRT) in the framework of the probabilistic test theory. For that real test results of former first year students in engineering science had been analyzed. The prototype of such a CAT has been developed. It focusses on a physics test for prospective students in the STEM fields. In fall 2021 the pilot phase was conducted with first year students in engineering science. The CAT shows that the same precision can be achieved with a mean of 9.3 items compared to 12 in the static test. The acceptance among the students is high. The correlation between the static test and the CAT is satisfactory.

1 INTRODUCTION

1.1 Starting Point

At the transition from school to university the knowledge level of prospective students in STEM – which stands for the academic disciplines “science, technology, engineering and mathematics” – is very heterogenous. In Germany about one-third of all STEM students (one in two in mathematics) drop out of university or change the subject before graduation [1]. The lack of basic knowledge is one of the most important reasons for dropouts. As one of the primary goals of current German education policy is to increase the number of graduates in the STEM fields, not only the number of beginners has to be increased but also the number of dropouts has to be reduced. As the first year at university is crucial, support offers should start as early as possible guaranteeing a smooth start.

1.2 The Project MINTFIT Hamburg

In 2014 the project “MINTFIT Hamburg” [2, 3] started to give support with a free-of-charge online service. The MINTFIT platform aims at prospective students who are interested in independently reviewing their knowledge in mathematics, physics, chemistry and computer science. One can check if the current level of knowledge fits the basic requirements of a STEM study in Germany and at the same time receive information about which gaps need to be filled. MINTFIT focuses specifically on the transition from school to higher education, since there are good opportunities to identify and fill knowledge gaps before problems occur in the first year. Within the MINTFIT learning platform an online test is always the starting point. That far all tests have a static character. For three tests (physics, chemistry and computer science) the questions do not have any variation whereas the math test randomly chooses questions from an item pool or varies values within one test item.

1.3 The Static MINTFIT Physics Test

For the CAT development, described in the following, the physics test was chosen from all MINTFIT tests. The static MINTFIT physics test, that is the current standard, consists of 40 basic test items in the subdomains mechanics, electricity, energy, and optics. While the contents correspond to the intermediate school level, the (re)presentation as well as the mathematical methods (e.g. vector analysis or integral calculus) are oriented to the university level and serve therefore as a well fitted preparation for higher education.

Based on the main topics and subtopics in school about 200 items had been developed. Different question types are used: multiple choice, drop down menu, cloze, short answer, and algebraic questions. For the final test 40 out of these 200 items have been selected to accommodate the test total length to about 60 minutes. The selection was done in several iterations. The items were tested in target-group schools and with first-year students. The final test has been adjusted to allow an optimal assessment of the level of knowledge as well as to offer suitable learning recommendations.

For the comparison to the CAT development described in the following only the mechanics part of the static test was used.

1.4 Alternative Test Ideas for the MINTFIT Physics Test

As the physics test is voluntary, the duration time of 60 minutes is quite long and seems to be one of the reasons for the high dropout rate. It shows up that only one third of the started tests are finished regularly. In about 20 % of the finished tests users stop answering after the first 10 out of 40 questions. The aim for a new test type is therefore to significantly reduce the number of questions or the test duration while maintaining the quality of competence assessment at the same time. To accomplish this, MINTFIT experiments with so called computerized adaptive testing methods [4]. This approach resembles more an oral than a written exam and is related to the expectation to achieve a better or equivalent assessment of participants' competences with fewer questions.

2 METHODOLOGY COMPUTERIZED ADAPTIVE TESTING

Computerized adaptive testing (CAT) is a form of testing, where the next administered question (item) depends on the answer of the already administered items. If, for example, a student answers an item correctly, the next item could be more difficult. Here CAT is used within the framework of the item response theory (IRT) [5]. More precisely the Rasch model [6] is used. The Rasch model defines every item with a single parameter, the difficulty b, and also every student with a single parameter, the ability θ. The probability of answering an item correctly is then determined by:

$$P(x = 1|\theta) = \frac{\exp(\theta - \mathrm{b})}{1 + \exp(\theta - \mathrm{b})} \tag{1}$$

One of the advantages of this model is, that ability and difficulty are measured on the same scale and can therefore be compared. Note that in this model, if a person has the same ability as an item’s difficulty, the probability of solving this item is only 50%. The maximum of the slope is at exact this point, where θ and b are equal. The slope is of special interest. The higher the slope is, the better an item can differentiate at this point between higher ability and lower ability students. To make this point clearer, imagine an item with difficulty 0, and four students with abilities of -0.5, 0.5, 1.5, and 2.5. The Rasch model then yields the following probabilities of solving the item: 0.38, 0.62, 0.82 and 0.92. See how the differences between the students shrink: From 0.24 between the first two students, to 0.1 between the last two students. It can be seen therefore, that the most information can be gained from an item, when the student has the same ability as the item difficulty. CAT tries to make the most efficient test by re-estimating the student’s ability after every answer and presenting the best fitting item. There are different stopping criteria possible. A stopping criterion that relies on the precision of the measurement is used. This means, the test ends as soon as a certain precision is reached and a statement about the student’s ability can be made with sufficient certainty. The precision of a

test is also called the test information. It is defined as the sum of the item information. The item information for an item i is defined as the product of the probability of solving the item (p_i) and the probability of not solving the item (q_i). The standard error of estimation SE is then defined as the square root of the inverse of the test information, or as a formula:

$$SE = \frac{1}{\sqrt{\sum p_i q_i}} \quad (2)$$

As many people are more familiar with the term standard error, it will be used throughout this paper. Note, that the Rasch model is a rather simple model. More complex models, like the 2PL model [5], describe an item not just with a difficulty parameter b, but with a second parameter a that has an influence on the slope. This parameter is called the discrimination parameter. The Rasch model can be seen as a special case of the 2 PL model, where the discrimination parameter is fixed to 1 for every item.

3 CAT PROTOTYPE

3.1 Concept

For the CAT prototype an algorithm is implemented that is based on the 2 PL model considering the difficulty and the discrimination parameter for every item. For the pilot phase described here the discrimination power of all items was set to 1. This is equivalent to using the Rasch model.

Initially, a test user's ability is estimated with the value 0. A new question is always added to the test based on the current estimate. For the selection of the next question the item information of all unanswered questions is calculated. One of the questions with (almost) maximum (> 99% of the maximum value) item information is randomly chosen next. Questions are excluded from which it can already be predicted with a probability of 95% or more whether they will be answered correctly or incorrectly.

Always after a question has been answered, an estimate of the ability is calculated for the respective test user and the standard error is determined. For this purpose, the highest maximum a-posteriori probability is taken. This is done by multiplying the discrete standard normal distribution between the minimum and maximum possible ability (-3 and 3) by the respective probabilities that a person with that ability answers the question in that way. The maximum value is the estimated ability. The algorithm terminates if, after answering a question, the standard error falls below the specified threshold. The determined ability after the last question gives the score for the whole test. For better understanding and consistency with the other MINTFIT tests, this value is converted into a scale from 0 to 3 stars and is given as feedback to the test user.

3.2 Realisation

The adaptive test is implemented as a Moodle plugin [7]. The implemented plugin borrows some ideas from a publicly available Moodle plugin for adaptive tests [8], but is not based on that, as a more general approach is used for the ability estimate and the selection of the questions.

A test can be configured with multiple subdomains, but in our test run the CAT was limited to mechanics. For each subdomain a question pool must be specified and a minimum number of questions has to be configured. Each question in the question pool needs a corresponding question difficulty and discrimination power (here set to 1), which are specified using tags in Moodle. The termination threshold for the standard error has to be specified and a maximum number of questions is also set as an alternative termination criterion for the entire test.

From a test user's point of view, the CAT has the appearance of a regular Moodle test. Due to the adaptive concept, the questions have to be answered in the given order and it is not visible in advance how many questions the test will contain.

3.3 Question Parameters

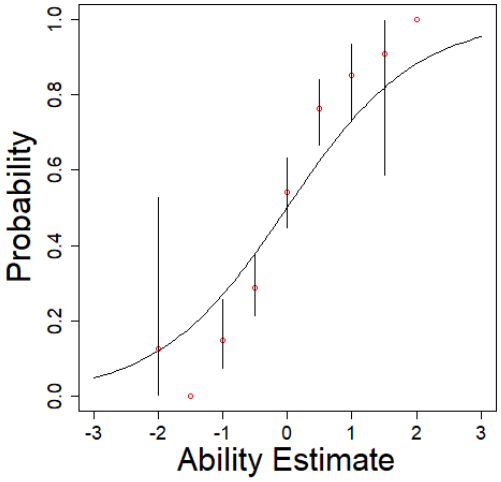

Fig. 1. Excellent fit to Rasch model

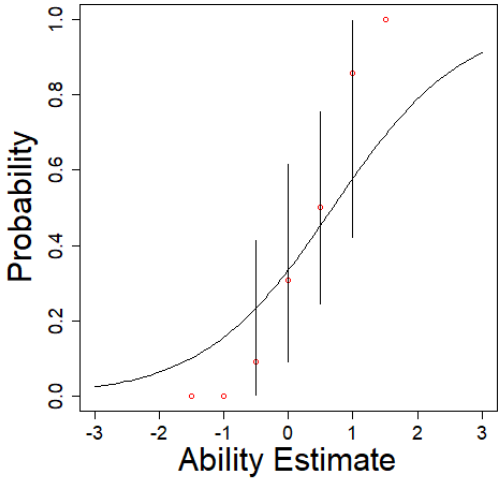

Fig. 2. Acceptable fit to Rasch model

As described above, for the physics test an item pool of 200 items had been developed. To determine the parameters as e.g. difficulty all items were investigated in the winter term 17/18 in a pretest with approx. 350 first-year students at the Hamburg University of Technology (TUHH) and the Universität Hamburg (UHH). The items were split into 8 test sets. 12 so-called anchor items (3 from each subdomain) were included in each test. Afterwards an IRT analysis of the items was performed. As the number of given answers for some items was quite limited the fit of the 2 PL model did not satisfactorily converge. Therefore, the discrimination power was fixed to 1 for all items and the fit was performed within the framework of the Rasch model. An important quality criterion for the procedure is the empiric match between data

and Rasch model. Two examples of the model agreement after the fit are shown in figures 1 and 2.

From the complete question pool 72 questions are in the mechanics subdomain and have been used in the CAT.

4 PILOT PHASE

The test setup for the pilot phase consisted of two tests, the mechanics part of the static MINTFIT physics test and the new CAT. Two questionnaires, one after each test, evaluate the respective tests. Also, in the second questionnaire personal information of the test user is included, e.g. the final grades in math and physics before university as these had been identified as important predictors for academic success [1, 9]. The idea is that each test user starts with the static test followed by the first questionnaire. Next is the CAT followed by the second questionnaire (see figure 3). The test feedback and example solutions for the CAT are given right after the test whereas the feedback of the static test is only available after the whole procedure not to give any support for solving the questions in the IRT-CAT test.

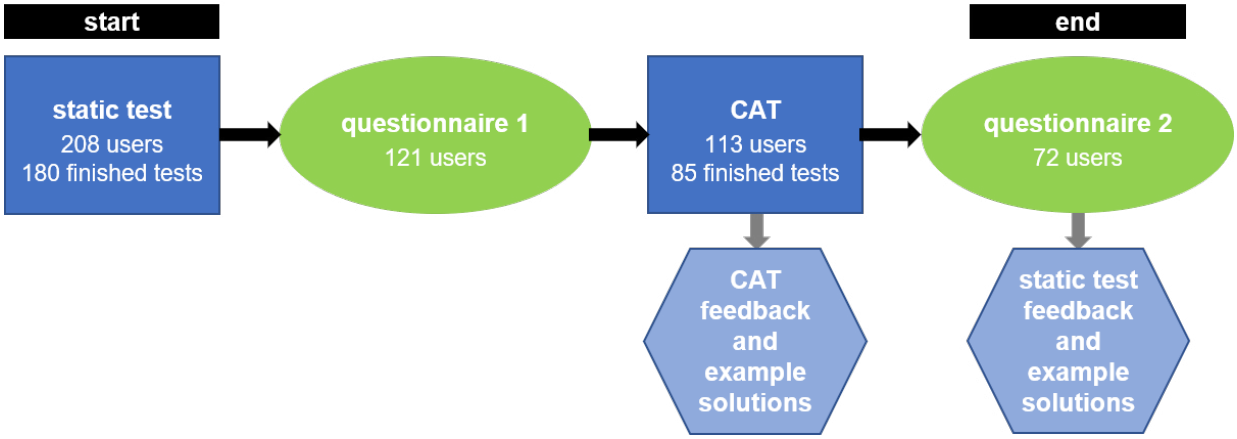

Fig. 3. Overview of test setup

In November 2021, 208 first year students in engineering science took part in the pilot phase. The students worked individually during a 90 minutes presence exercise session that complements the module in mechanics.

Out of the 208 test users that started the static test, 85 users got a converging result for the CAT. Finally, 72 test users also filled in the two questionnaires and thus completed all 4 stages.

To get an impression on how the IRT-CAT-test works in practice for a single user the development of the difficulty of the questions and the estimate of the personal ability and confidence intervals are shown in figure 4.

The items of the static test were not excluded from the item pool for the CAT. 70 % of the converged CAT runs had no overlap at all and 29 % of the tests had one

question from the static test included in the CAT. This has no impact on the analysis and will be neglected in the following.

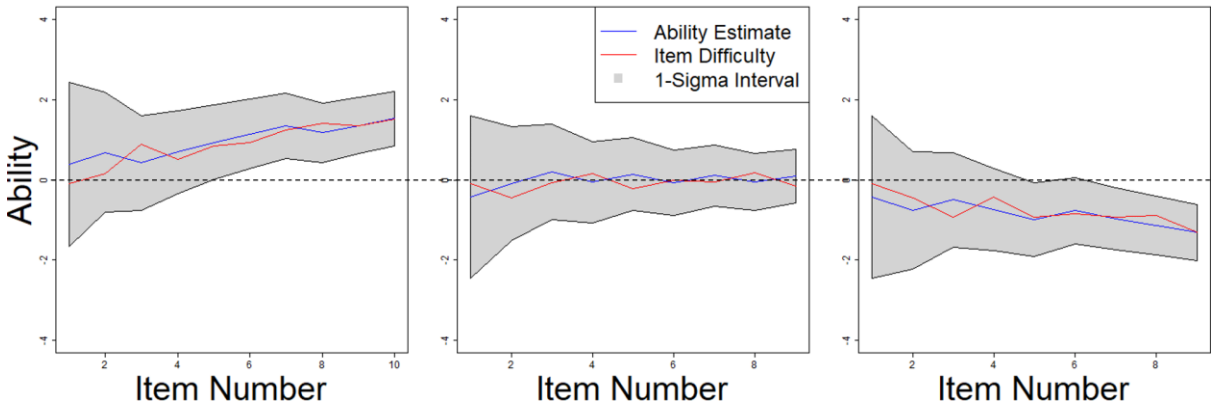

Fig. 4. Development of personal ability estimate and question difficulty versus item number for a test user with high (left hand), mean (middle) and low (right hand) personal ability

5 RESULTS

5.1 Comparison between CAT and Static Test

For the following analysis only the test results have been used where all four steps of the test procedure had been completed. As in the static test each user is forced to complete all 12 items, the number of items in the CAT depends on how many questions are necessary to fall below the error threshold. This leads to the distribution of the number of questions shown in figure 5. The mean value is 9.3 items. So, the number of items could be reduced by 23 % on average compared to the static test.

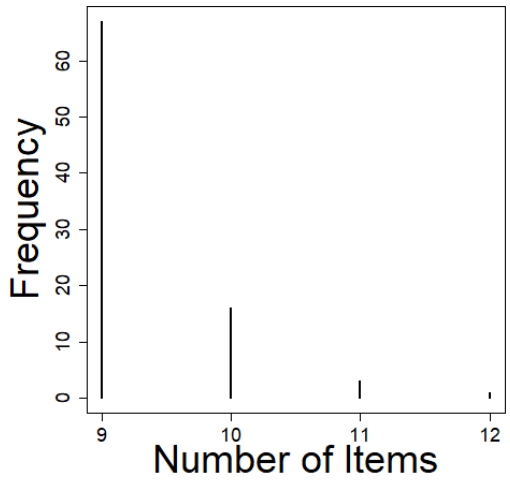

Fig. 5. Distribution of number of items in CAT runs

The comparison of the standard error of the personal ability versus the personal ability for the static test and the CAT is shown in figure 6. The error threshold for the

CAT is constant over the whole range of the personal ability whereas the error in the static test is a little bit lower around the mean value for the personal ability but rises to low or high values of the personal ability. Over the full range the error is comparable on average.

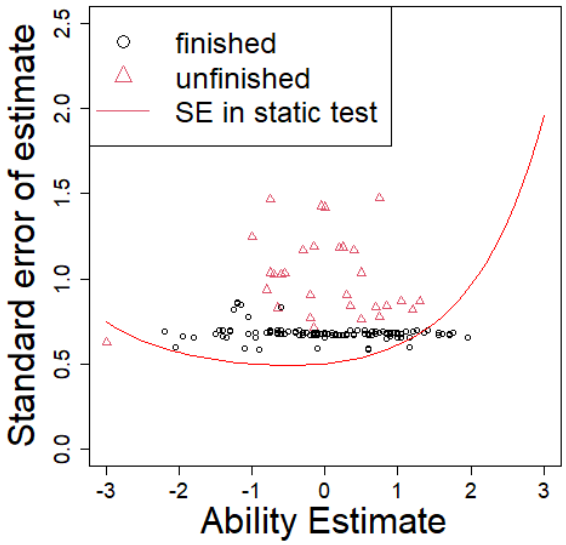

Fig. 6. Standard error versus personal ability estimate for static test (solid line) and CAT (finished and unfinished test runs)

As a quality check if the static and the CAT measure the same ability the correlation of the test results was determined. A satisfactory value of 0.56 was achieved (see table 1).

Table 1. Correlation between static test, CAT and school grades

	Static test	CAT	Grade in math	Grade in physics
Static test	1			
CAT	0.56	1		
Grade in math	0.46	0.49	1	
Grade in physics	0.56	0.35	0.47	1

For the correlation between the grades in math and physics a difference shows up for the two tests: the static test has a stronger correlation to the grade in physics whereas the CAT has a stronger correlation to the grade in math. From the social parameters only the academic background of the parents leads to a significant difference between the tests. The correlation between the test result and that at least one of the parents has an academic background is 0.12 for the static test and 0.31 for the CAT. Together with the difference in the correlation of the school grades this

might be a hint that the two tests do not measure exactly the same ability of the test user. For a detailed understanding further investigation is needed.

5.2 Evaluation Results of CAT

In addition to the outcome of the test results the acceptance and judgement of the test users was determined with the two questionnaires.

In good agreement with the reduction of the average number of questions the perception of the duration of the test was significantly shorter for the CAT than the static test. On a five-point Likert scale ranging from “very short” to “very long” 45% chose short and 6 % very short for the CAT whereas only 25% selected short and none very short for the static test.

Both the static test and the CAT were rated in general as good by the majority: 58% “very good” and “good” for the CAT, 63% “very good” and “good” for the static test on a five-point Likert scale ranging from “very good” to “very bad”.

Astonishingly also the difficulty of the questions was comparably evaluated. In the CAT, 58% of the test takers rated the difficulty of the questions as “average”, and also 58% did this for the static test. The expectation was that the CAT mostly shows a mean rating as the questions were selected to fit the personal ability whereas the static test always has questions over the whole difficulty range, so weak students should give a feedback as the questions being “difficult” and strong students have the impression that the questions are “easy”.

6 SUMMARY

The prototype of the CAT was successfully implemented and tested. The number of questions could be reduced significantly by 23 % within a comparable error and hereby fulfilled the expectations. The assessment of the test users for the length of the CAT compared to the static test corresponds to the measured reduction. A satisfactory correlation between the results of the CAT and those of the static test was achieved. As the detailed knowledge of the item parameters is crucial, a further reduction of the test length seems achievable when there is more data available and the discrimination power of the items can also be determined.

7 ACKNOWLEDGMENTS

MINTFIT Hamburg is a joint project of the Hamburg University of Applied Sciences (HAW), HafenCity University Hamburg (HCU), Hamburg University of Technology (TUHH), University Medical Center Hamburg-Eppendorf (UKE) as well as Universität Hamburg (UHH) and is funded by the Hamburg Authority for Science, Research, Equality and Districts (BWFGB).

REFERENCES

[1] Heublein, U., Ebert, J., Hutzsch, C., Isleib, S., König, R., Richter, J. and Woisch, A. (2017), Zwischen Studienerwartungen und Studienwirklichkeit, Ursachen des Studienabbruchs, beruflicher Verbleib der

Studienabbrecherinnen und Studienabbrecher und Entwicklung der Studienabbruchquote an deutschen Hochschulen, *Forum Hochschule,* 1|2017, DZHW, Hannover.

[2] Schramm, T. (2015), Mintstudium Hamburg – Eine konzertierte Aktion. Proc. 12. Workshop Mathematik für Ingenieure, HCU Hamburg 2015, Frege-Reihe Hochschule Wismar.

[3] Sitzmann, D. (2020), Fit ins Studium mit MINTFIT Hamburg – Unterstützungsangebote für ein erfolgreiches MINT-Studium und zur Senkung der Studienabbruchquote. So gelingt E-Learning – Reader zum Higher Education Summit 2019; Studienergebnisse und Praxisberichte zum Einsatz von E-Learning an deutschsprachigen Hochschulen. Stephan Kahmann, Prof. Dr. Stefan Ludwigs, Pearson, München.

[4] Magis, D., Yan, D., and Davier, A. A. v. (2017), Computerized Adaptive and Multistage Testing with R. Springer.

[5] Hambleton, R.K. & Swaminathan, H. (1985), Item Response Theory: Principles and Applications, Kluwer-Nijhoff.

[6] Rasch, G. (1960), Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research, Copenhagen 1960, expanded edition with foreword and afterword by B.D. Wright. The University of Chicago Press, Chicago 1980.

[7] (2022, April 28), Moodle Homepage, https://moodle.org.

[8] (2022, April 28), Moodle plugin: Adaptive Quiz, https://moodle.org/plugins/mod_adaptivequiz.

[9] Sorge, S., Petersen, S., Neumann, K. (2016), Die Bedeutung der Studierfähigkeit für den Studienerfolg im 1. Semester in Physik, *Zeitschrift für Didaktik der Naturwissenschaften* 22,1, pp. 165-180.