Turning CAT into MOUSE: Adaptive Testing Mechanisms in Student Exercise Selection

Ulrike Padó[1], Konstanze Mehmedovski[1], Anselm Knebusch[1]

[1]Hochschule für Technik, Schellingstr. 24, Stuttgart, 70174, Germany.

Contributing authors: ulrike.pado@hft-stuttgart.de;
konstanze.mehmedovski@hft-stuttgart.de;
anselm.knebusch@hft-stuttgart.de;

Abstract

Computerised Adaptive Testing (CAT), traditionally employed in high-stakes summative testing, can also offer significant potential for enhancing formative self-assessment: In this paper, we use it to adapt exercise question difficulty based on student responses, ensuring questions are appropriately challenging and thereby supporting students' progression from lower ability levels to greater competency. We present a software tool and extensive usage recommendations in order to facilitate the use of adaptive self-assessments and exercises for educators and students, with the overall goal of creating inclusive and effective learning experiences in diverse student populations in Higher Education. This article presents the modification of an existing Moodle plugin for formative CAT, emphasising informative feedback for iterative learning and pragmatic solutions for integrating existing question pools – thus turning CAT into MOUSE (Module Optimised for Usability in Student Exercises).

Keywords: Computerised Adaptive Testing, Moodle, Exercises, Formative Assessment

1 Introduction

In Higher Education, ”one size” of teaching and exercises rarely ”fits all” students: Especially in the starting semesters and outside the research university context, students have heterogeneous previous knowledge from their previous educational backgrounds (see Bressoud et al. 2013 for the case of Mathematics). Educators therefore look for ways to adapt their content to their students' individual needs, for example by providing additional exercise sheets with remedial or advanced questions. Ideally, they should be supported by technology in this task.

A proven method of sampling comparable, yet different questions at varying difficulty levels is Computerised Adaptive Testing (CAT, see Wright 1988). CAT techniques autonomously adapt question selection under uncertainty with the goal of increasing estimation accuracy by taking a student's already observed performance into account. Although the algorithm is very dissimilar from current trends in AI, CAT therefore arguably qualifies as an *intelligent actor* that ”takes the best possible action in a situation”, the definition of an AI system in Russell and Norvig (2016). In contrast to machine learning approaches like, e.g., Large Language Models, it does so transparently and without recourse to large amounts of training data (once question difficulty is known or estimated – we explore different strategies of obtaining question difficulty estimates in Section 4).

CAT was developed to allow efficient high-stakes summative assessments. Here, we explore the approach for formative self-assessments and exercise sheets that focus on fostering individual learning and development. Our driving question is how to facilitate the use of CAT for formative adaptive tests that benefit students and and easy to prepare for teachers. This situation can be analysed using Knezek and Christensen's Will, Skill, Tool and Pedagogy (WSTP) dimensions (Knezek and Christensen 2015). This framework identifies four main predictors of school educators' technology integration in the classroom: Will is educators' attitude to the use of technology in teaching, Skill is their confidence and ability to do so, Tool their access to the necessary tools and Pedagogy their confidence that their use of technology will benefit their students. Framed in these dimensions, we present a modified CAT Tool as a technical solution that addresses heterogeneity in students' abilities. We also show how we worked to lower the Skill level needed to use it. This is a significant concern, as Lilley et al. (2011) find that educators' top reason to not use CAT is that it is ”too complicated to implement”, followed by the need of a calibrated question pool, which we will also address. We also contribute along the Pedagogy dimension, giving insights gained during development and testing of our tool with the goal of increasing educators' confidence in their use of the technology, and enabling them to actively support and guide their students' individual learning process.

Our Tool is based on the widely-used open-source learning management system (LMS) Moodle[1], and more specifically its Adaptive Quiz plugin[2] that supports the straightforward use of CAT in testing. However, the originally summative philosophy of the plugin requires adjustments to make it more suited to formative testing and exercise selection and that reduce the Skill level required for using it. By this, we turn summative CAT into formative MOUSE - a Module Optimised for Usability in Student Exercises [3].

Our contribution in this paper is (a) the development and description of an easy-to-use Tool for formative adaptive tests and exercises, (b) the introduction of a well-structured process for teachers to follow as they implement these tests for improving

[1] https://www.moodle.org

[2] https://moodle.org/plugins/mod_adaptivequiz

[3] The code for the MOUSE plugin is published as Open Source software at https://transfer.hft-stuttgart.de/gitlab/knight/adaptivetests

their Skill, and (c) best practices, examples and suggestions for each step of the process, based on our pilot studies, to help optimise the outcome for students and teachers according to the Pedagogy dimension.

The paper is structured as follows: We first discuss CAT and the adjustments necessary for its use in a formative context, also referencing previous work (Section 2). Section 3 outlines the process of using MOUSE and gives insights into the educator's role and possible challenges at each step. We structure this part by our Tool and Pedagogy contributions (i.e., technical adaptations and recommendations for classroom use). Sections 4 and 5 present practical usage examples addressing question pool creation and a field test of the MOUSE process with qualitative and quantitative evaluation; we close with a discussion in Section 6.

2 Adaptive Testing and Formative Tests

Computer-adaptive testing (Wright 1988) is based on Item Response Theory (IRT, Lord 1980, see also van der Linden 2010), which establishes a joint scale of measurement of item (question) difficulty and test-taker ability such that students who attempt a question that matches their ability level have a 50% chance of answering correctly, and a higher chance of answering questions that are less difficult than their ability level. CAT estimates a student's ability level with increasing accuracy by selecting a string of questions at or near the student's current presumed ability, observing the outcome and adapting the ability estimate and question choices accordingly.

As a result, high-performing test-takers encounter fewer overly simple questions, reducing errors from inattentiveness or over-thinking, while lower-performing individuals face fewer overly challenging questions, decreasing the likelihood of random guessing. This flexibility makes adaptive testing particularly attractive for classrooms with a wide range of ability levels and for formative testing or exercise selection.

In CAT tests, students' test-taking experience differs from static tests containing a wide range of question difficulties: With CAT, not all students receive the same questions – just like in oral exams, this is not absolutely necessary for fair testing. Additionally, participants can expect to only solve about half of the presented questions on average, regardless of their effort or performance level, because the algorithm matches the questions to their current estimated IRT ability level, where they have a 50% chance of failure. (They will also be presented with some questions above or below their level, where their chance of success is accordingly lower or higher, further contributing to the average 50% success rate.) As examined in Frey et al. (2009), high-performing test-takers may initially be frustrated by a perceived low question success rate, and lower-performing students surprised by a poor overall test score, despite having solved an unusually high number of questions. These differing characteristics introduce the ethical obligation to inform students of the specifics of the CAT process beforehand to allow them to choose appropriate strategies and avoid frustration.

However, we ultimately consider these characteristics beneficial for our use case, since they ensure that students at all levels efficiently focus on questions of appropriate difficulty. Students at lower levels can repeat the exercise sheet with new questions until they reach the educator's goal level; students at higher levels are appropriately challenged and profit from the time they invest.

A second distinction from conventional tests is the way results are presented. Adaptive tests report performance as the achieved level rather than the typical percentage or grade. This numerical value may not be intuitively interpretable for the students. In formative assessment, more detailed feedback should therefore be provided to guide students on their learning journey.

Beyond the Adaptive Quiz plugin to the LMS Moodle, several more software tools exist that implement summative CAT for Higher Education, both developed in an academic context like Concerto (Harrison et al. 2020), AdaptiveTesting (Oppl et al. 2017) and KAT-HS (Fink et al. 2021), and available commercially like FastTest[4].

Closer to our goals of creating adaptive exercise sheets for learning, using the CAT approach for formative testing is also well-established: McCallum and Milner (2021) positively evaluated a CAT approach where students felt that repeated formative CAT testing helped them monitor their progress and encouraged them to study, and Yang et al. (2022) proposed to add a review function for past content. However, we have not found an easy-to-use, freely accessible tool for formative CAT testing. We therefore chose the Moodle Adaptive Quiz plugin as the basis for our concrete implementation of formative adaptive CAT, seamlessly integrating the tool into a popular LMS familiar to academic users worldwide. Our learnings of course equally apply to any other implementation.

3 The MOUSE Process

In the following, we introduce the MOUSE process of preparing and using CAT for formative exercises (Fig. 1) to describe both our adaptations to the plugin (marked by **Tool**) as well as our recommendations to educators (marked as **Pedagogy**). These recommendations stem from the multi-year development of the MOUSE process, which implemented formative CAT in three undergraduate courses at our University. During development, the key request from the students' side was to be shown feedback on their individual answers in addition to receiving an ability level, so they could review any mistakes. On the educators' side, we found a need to support the pragmatic use of existing partial-credit questions and to make recommendations for issues relating to the question pool, like pool size, manual question difficulty assignment and quality control on difficulty estimates (Padó et al. 2025).

The text follows the five major steps in the MOUSE process and discusses each aspect of the steps in turn. Sections 4 and 5 contribute real-life usage examples.

3.1 Scenarios and Desired Feedback

The first step of the MOUSE process is to consider the fit of the method with the content and goals of an educator's class. This includes choosing the right usage scenario, deciding on the number and characteristics of the difficulty levels that will be used by the underlying CAT algorithm and, finally, determining the passing level based on the

[4] https://assess.com/computerized-adaptive-testing/

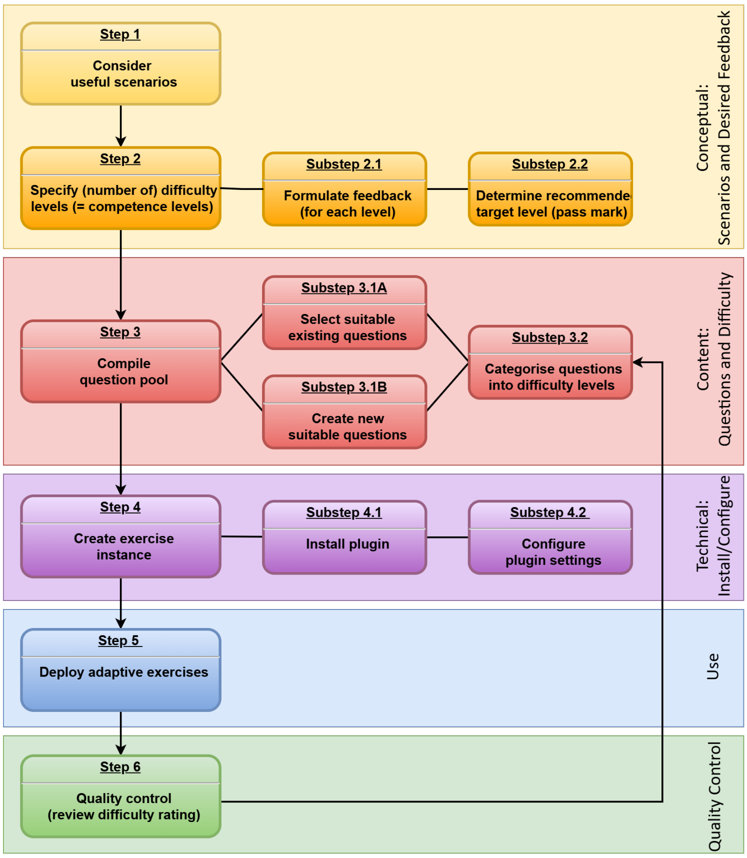

Step 1
Consider useful scenarios
Step 2
Specify (number of) difficulty levels (= competence levels)
Substep 2.1
Formulate feedback (for each level)
Substep 2.2
Determine recommende target level (pass mark)
Conceptual: Scenarios and Desired Feedback
Step 3
Compile question pool
Substep 3.1A
Select suitable existing questions
Substep 3.1B
Create new suitable questions
Substep 3.2
Categorise questions into difficulty levels
Content: Questions and Difficulty
Step 4
Create exercise instance
Substep 4.1
Install plugin
Substep 4.2
Configure plugin settings
Technical: Install/Configure
Step 5
Deploy adaptive exercises
Use
Step 6
Quality control (review difficulty rating)
Quality Control

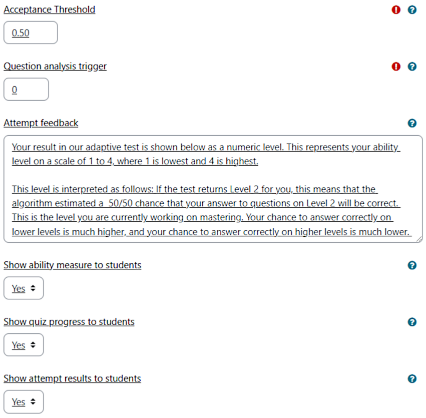
Fig. 1: MOUSE process

skills that the educator wants the students to master easily. Understanding the characteristics of the different difficulty levels also makes it easier to formulate feedback to students that helps them interpret their test results and take appropriate next steps.

Scenarios The first step to maximise overall student benefit in adaptive exercises is from the realm of **Pedagogy**. In our experience, adaptive tests worked well in formative self-assessment scenarios and practice exercises that helped students prepare for their summative exams (Padó et al. 2025). Additionally, since these situations are low-stakes, the requirements on the question pool are lower than for high-stakes, large-volume testing and are realistic to fulfil even for small cohorts (see Section 3.2).

We used short adaptive tests with relatively homogeneous content in our formative setting (”one test per topic”). This is a simple measure to avoid issues where test-takers may not be presented with questions from all topic areas for tests covering multiple topics. Additionally, students received feedback early on and were able to purposefully access exercises targeting the specific competencies they wanted to work on.

Specify difficulty levels For each chosen scenario, the number of difficulty levels specified for the MOUSE plugin determines how fine-grained the algorithm’s estimates



Acceptance Threshold

0.50

Question analysis trigger

0

Attempt feedback

Your result in our adaptive test is shown below as a numeric level. This represents your ability level on a scale of 1 to 4, where 1 is lowest and 4 is highest.

This level is interpreted as follows: If the test returns Level 2 for you, this means that the algorithm estimated a 50/50 chance that your answer to questions on Level 2 will be correct. This is the level you are currently working on mastering. Your chance to answer correctly on lower levels is much higher, and your chance to answer correctly on higher levels is much lower.

Show ability measure to students

Yes

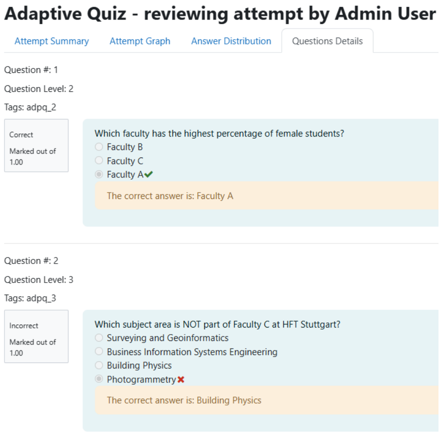
Show quiz progress to students

Yes

Show attempt results to students

Yes

Fig. 2: Plugin settings for our modifications (top: partial credit and trigger for quality control reminder, bottom: individual question feedback), showing also some sample informative feedback on the ability levels.



Adaptive Quiz - reviewing attempt by Admin User

Question #: 1

Question Level: 2

Tags: adpq_2

Correct

Marked out of 1.00

Which faculty has the highest percentage of female students?

○ Faculty B

○ Faculty C

● Faculty A ✓

The correct answer is: Faculty A

Question #: 2

Question Level: 3

Tags: adpq_3

Incorrect

Marked out of 1.00

Which subject area is NOT part of Faculty C at HFT Stuttgart?

○ Surveying and Geoinformatics

○ Business Information Systems Engineering

○ Building Physics

● Photogrammetry ✗

The correct answer is: Building Physics

Fig. 3: Student's review of an adaptive test attempt using the standard Moodle view and showing dummy questions. Attempt review can be allowed using the bottom setting in Fig. 2.

and therefore the feedback to students will be. We recommend starting off with 4–5 difficulty levels. On the one hand, this provides sufficient detail for feedback, and on the other, it ensures the question pool size remains manageable (see Section 3.2) and question difficulty levels are clearly distinguishable. Each difficulty level should be characterised by specific learning objectives or sub-skills, representing successive competence levels. This facilitates both question classification and the creation of meaningful feedback for the students.

Formulate Feedback on Ability Level Conventionally, grade categories are used to assess performance, e.g., ranging from 1 to 5 or A to D. In contrast, adaptive tests provide an estimate of the student's ability level (along with an associated estimation error). To make this test result meaningful to students and support their learning, it requires proper interpretation. From the level descriptions for the required skills (learning objectives) created during the categorisation of task difficulty, a clear, verbalised guide can be easily developed to help students understand the meaning of their reported ability level.

Our experience in Padó et al. (2025) shows that students wish to receive even more fine-grained feedback by reviewing incorrect responses and their correct answers. On the **Tool** level, we therefore included a new option in the adaptive quiz instance settings, allowing educators to enable students to use a standard Moodle view to review their own attempts (see bottom of Fig. 2 and Fig. 3).

Determine pass mark If a student is assigned a CAT ability level, they have not mastered this level yet, but have a 50% chance of failing each question and should continue practising at this difficulty. The pass mark should therefore be at least one level above the minimum requirements for the activity.

3.2 Content: Questions and their Difficulty

The next step in the MOUSE process involves the preparation of the questions for each adaptive test or exercise sheet. We give recommendations on the size of the question pool and discuss issues like appropriate Moodle question types and the handling of partial-credit questions.

The crucial task is to generate a question pool with an appropriate number of questions on the different difficulty levels. Normally, difficulty is established in norming studies with large numbers of participants (Downey 2010), but this is often prohibitive for smaller learning cohorts. In the easiest case, existing questions with success frequency statistics can be re-used from earlier tests; reliably assigning difficulty to existing or new questions without usage statistics is more complex and time-consuming. We give general recommendations here on how to proceed and report on our experience with a real-world question pool in Section 4.

Question Pool

We first turn to **Tool** level questions of question pool size, question format and the handling of partial credit.

Question pool size To ensure smooth functionality and prevent the algorithm from terminating abruptly, it is essential at the **Tool** level to maintain a sufficient number of questions across all difficulty levels. The exact number of questions needed depends on the planned test length. In our experience, providing 7–10 questions per level was sufficient for tests consisting of 8–15 total questions. Educators should monitor individual test results regularly to address a lack of questions if it occurs.

Question format The adaptive algorithm introduces the **Tool**-level constraint of question format because it requires auto-gradable question types (closed-format question types, like Moodle's single-choice, multiple-selection, and matching items). Additionally, (semi-)open format questions can be included if they are reliably automatically evaluated, for example, using a computer algebra system like STACK[5] for appropriate questions in Mathematics.

Question credit The CAT algorithm in the MOUSE plugin expects all questions to be graded as correct or wrong, while existing question pools may also contain questions that award partial credit. The original Adaptive Quiz plugin treats all partially correct answers as correctly solved, which is clearly over-optimistic in cases like 0.25 achieved credits out of 1 total. This design decision at the **Tool** level reduces the number of existing questions that can be re-used. The MOUSE plugin therefore offers an option to flexibly set a correctness threshold so educators can define the percentage of total points (e.g., 0.5 or 0.75) required to consider a partial-credit question as solved correctly. This allows the re-use of existing partial-credit questions.

Question difficulty categorisation

One of the focal points of implementing adaptive exercises is at the intersection of the **Tool** and **Pedagogy** levels: Determining the difficulty of the questions to be used. In our experience, extensive norming is impossible for many educators, especially for

[5] https://stack-assessment.org/

smaller or infrequently taught courses, so we recommend pragmatic approaches to estimating difficulty in case normed levels are not available.

Existing Questions with Usage Statistics Creating a MOUSE question pool is easiest if questions with previous usage statistics exist. In this case, the frequency of correct answers from previous conventional test deployments (while not sufficient to create an IRT difficulty estimate) is an initial basis for assigning appropriate levels, for example by defining frequency cutoffs for the desired amount of levels. Often, this analysis shows that particular levels are over- or underrepresented. Overrepresented levels can be reduced by sampling, but a larger question pool is no disadvantage to the CAT algorithm. Underrepresented question levels should be supplemented by new questions of similar difficulty. See Section 4 for a real-world example.

Existing Questions without Usage Statistics require manual assignment of difficulty estimates. This is not an easy task – Hamamoto Filho et al. (2020) find that experts' estimates of question difficulty only match the real difficulty 54% of the time, with more over-estimation (assigning a too-high difficulty level) than under-estimation.

We experimented with different approaches and present more in-depth results in Section 4. Our recommendation is to assign existing questions a difficulty level based on the educator's experience and to fill in missing levels by using Scaffolding (Bruner 1978). In this approach, questions are progressively simplified (i.e., prepared for lower difficulty levels) by providing hints or additional instructions to guide the test-taker. Inversely, existing hints may be dropped to make a question harder (see Ueno and Miyazawa 2018 for a scaffolding approach in an IRT context). Question difficulty assignments should be reviewed and corrected after using the questions in practice (see Section 3.5).

New questions When creating new questions, there are positive examples in the literature for using Bloom's taxonomy of cognitive processes (Anderson and Krathwohl 2014). Tiemeier et al. (2011) and Kim et al. (2012) find a significant correlation between the targeted Bloom's levels and observed question difficulty for questions written to explicitly cover different taxonomy levels. We recommend this approach for the creation of new question sets or the targeted extension of question sets for underrepresented difficulty levels. The scaffolding approach described above can be used in addition to Bloom's taxonomy levels or by itself. Again, assigned levels should be verified against student performance in the MOUSE quality control step 3.5.

3.3 Install and Configure

This step is purely on the **Tool** level: The installation of the MOUSE plugin and setting up of adaptive exercises follows typical Moodle procedures. Plugin settings can be adjusted to reflect the parameters decided in the previous steps, like the correct Moodle question pool, the minimum and maximum difficulty levels, etc. All other settings can be adjusted freely at a later time (see Fig. 2).

3.4 Use

Deployment finally brings the student in touch with the MOUSE tests. Since our **Tool** is a plugin to the students' familiar LMS, they have no trouble navigating the tests, in our experience. From a **Pedagogy** point of view, we recommend to highlight in advance two differences to static tests in Moodle: Students cannot skip questions and return to them later due to the context dependence of the question selection mechanism. Also, they can expect to only solve about half of the presented questions on average, but this rate does not indicate ability.

Additionally, students should know when and to which extent they will receive feedback (for example, immediately after the test, both at the test level and for performance on individual questions).

3.5 Quality Control

Regular reviews of question categorisation are essential to ensure consistency and validity of the difficulty levels. To facilitate such reviews, we have incorporated an automated notification system on the **Tool** level that prompts educators to periodically revisit and validate their question classifications using the Adaptive Quiz plugin's existing question analysis feature (see Section 5 for an example analysis).

4 Question Difficulty: A Real-World Example

Our recommendations for question difficulty estimation for existing questions merit some empirical corroboration. Over the course of the project, we experimented with a variety of approaches. We worked on a question set that is, based on our experience, typical in size of existing question sets in non-standardised testing for cohorts of 40-60 students. Questions were drawn from two different classes. The first subset consisted of 27 closed-format questions from several previous exams in an Introduction to Programming course (that contained also open-format questions, which we ignored), along with the observed frequency of correct answers for evaluation of the approach. A second subset, from the corresponding Advanced Topics in Programming course, contained 49 questions.

We first demonstrate the distribution of the observed success frequencies for our larger data set, then go on to discussing (failed) attempts at predicting question difficulty through machine learning and finally present a small study on assigning Bloom's taxonomy levels to existing questions. Experiences with Scaffolding for mathematical (STACK) questions conclude.

4.1 Observed Success Frequencies

Figure 4 shows the distribution of observed success frequencies (dichotomised from partial credit where needed using a 50% acceptance threshold). For example, the two questions on the bottom left were answered correctly just under 20% of the time.

Boxes mark the different difficulty levels that we assigned. When determining the levels, our goals were to define frequency corridors of (mostly) equal width that fit with observable frequency "steps" in the data (as between levels 3, 4 and 5), while respecting the observed maximum difficulty in order to not affect the overall difficulty level of the test by including harder items. In Fig. 4, this meant classifying every question with less than 50% success frequency as (the hardest) level 5, and moving

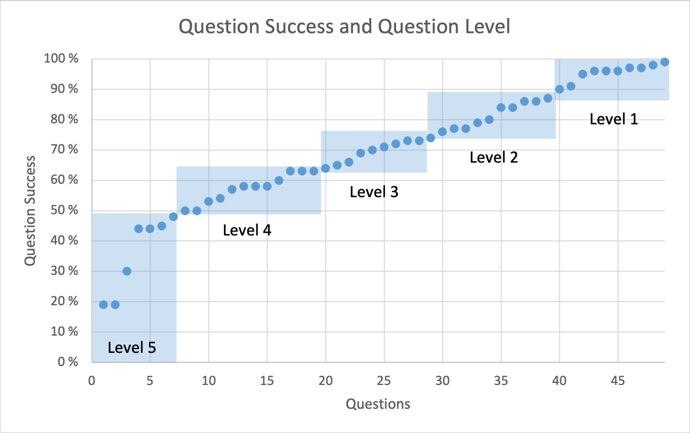


Fig. 4: Success frequency distribution and question difficulty levels (49 questions)

on in steps of 10-12 percentage points thereafter. The frequency distribution suggests that a sixth level could be added for the two to three hardest (lowest) items. In this case, more items at level 5 and 6 would need to be added.

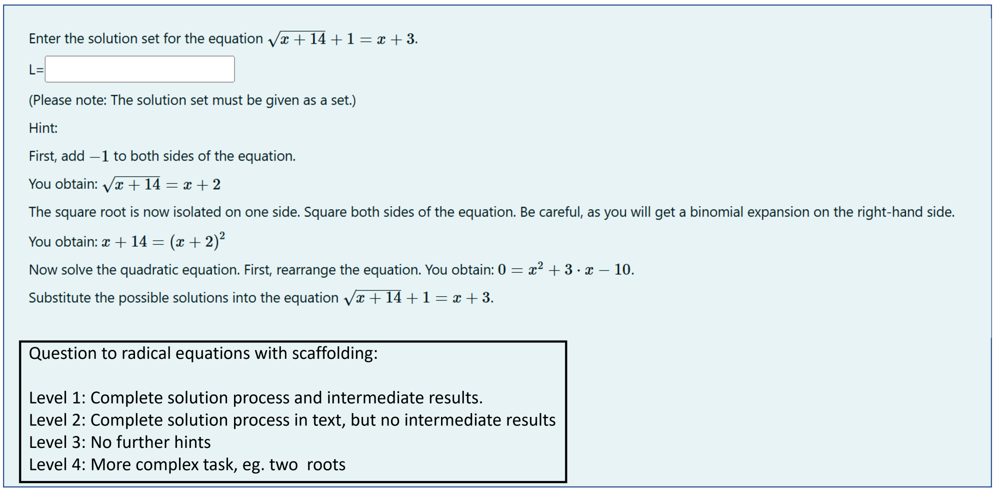
In Fig. 4, the chosen frequency bands do not clearly distinguish between all levels, so students' result patterns on these questions were monitored in the Quality Control step of the MOUSE process and levels adjusted accordingly.

4.2 Predicting Difficulty for Existing Questions

Predicting question difficulty for existing questions without success frequencies can plausibly be approached with machine learning algorithms. These promise objectivity in level assignment and significant reduction in manual work. Additionally, we explored two theory-driven methods: Categorising questions using a cognitivist approach and a variant of Bloom's taxonomy of cognitive processes (Anderson and Krathwohl 2014), the other uses the constructivist approach of Scaffolding (Bruner 1978). For our small study, we selected a subset of 34 questions from the two data sets introduced above (27 and 49 questions, respectively), such that the questions spread well across the observed success frequency levels.

Machine Learning (ML) For classical supervised ML, test and training data need to be annotated with the target difficulty levels as well as features that numerically describe relevant properties of the data; training and test sets should number at least hundreds of items. Since our real-life data sets are much smaller, classical ML approaches are generally not very promising for our question pool, and we were indeed not successful predicting question difficulty from features like question length, number of answer options and percentage of correct answer options (out of all options) for the 34 multiple-choice questions.

If no sufficient training data is available, using pre-trained Large Language Models without further fine-tuning seems like an interesting avenue to pursue. Raina and



Enter the solution set for the equation $\sqrt{x+14}+1=x+3$.

L=

(Please note: The solution set must be given as a set.)

Hint:

First, add -1 to both sides of the equation.

You obtain: $\sqrt{x+14}=x+2$

The square root is now isolated on one side. Square both sides of the equation. Be careful, as you will get a binomial expansion on the right-hand side.

You obtain: $x+14=(x+2)^2$

Now solve the quadratic equation. First, rearrange the equation. You obtain: $0=x^2+3\cdot x-10$.

Substitute the possible solutions into the equation $\sqrt{x+14}+1=x+3$.

Question to radical equations with scaffolding:

Level 1: Complete solution process and intermediate results.
Level 2: Complete solution process in text, but no intermediate results
Level 3: No further hints
Level 4: More complex task, eg. two roots

Fig. 5: Example question: Scaffolding through hints. Shown: Level 1, complete solution process with intermediate results; hints are reduced for harder questions (see inset).

Gales (2024) report that LLM zero-shot prompting performs best for ranking multiple-choice questions by difficulty, but their approach still only reaches a Spearman's ρ correlation coefficient of 0.4, showing that the task is far from trivial.

Bloom's Taxonomy Bloom's taxonomy (Anderson and Krathwohl 2014) describes the cognitive process needed to work on different (examination) classes of tasks. By assigning these classes, we hope to gain a an approximate level prediction for existing questions. We recruited three instructors who had previously taught the corresponding classes and asked them to assign Bloom cognitive process levels to the 34 questions. Annotation agreement was low (inter-rater reliability measure Krippendorff's $\alpha < 0.66$). We therefore continued with the more stable level assigned by the majority of graders. However, these assignments did not correlate with the observed success frequency at all (Spearman's $\rho = -0.1$, ns). Our results align with Hamamoto Filho et al. (2020) who also classify existing multiple-choice questions and find numeric, but no significant differences between the levels of taxonomy regarding difficulty and find that experts' estimates of question difficulty only match the real difficulty 54% of the time. Kibble and Johnson (2011) report similar results. Recall that literature findings are much more encouraging for the Bloom-based targeted creation of new questions in contrast to the classification of existing questions attempted here.

Scaffolding Scaffolding in this case takes the shape of hints or additional information that makes the question easier to solve. This approach was especially efficient in practice for STACK questions in Mathematics. STACK allows for the generation of randomised tasks with parameters by leveraging a computer algebra system. Therefore, it was sufficient to add scaffolding to the question templates in order to generate potentially unlimited question variants at the desired level. Figure 5 shows an example question template and scaffolding for different difficulty levels. At the lowest level (shown), the process of solving the radical equation is spelt out with intermediate results, leaving only the final steps to the student. The second difficulty level would show the process, but no intermediate results, and at level 3, no hints would be given.

For one test, we created 20 new question templates drawing inspiration from an existing extensive pool of approximately 1,000 STACK questions for use in foundational mathematics courses. From each template, we generated variants corresponding to four or five difficulty levels by adding or removing scaffolding hints; this was about 50% faster than generating 80-100 different questions from scratch. Additionally, the difficulty levels were clearly distinct, could be described unambiguously for feedback creation, and were comparable between templates.

5 MOUSE Field Test

To conclude, we present a real-life instance of the full MOUSE process. We set up adaptive exercises in a Mathematics class at our University and evaluated student satisfaction and performance qualitatively and quantitatively (see Padó et al. 2025).

To pass the class, students took mandatory (non-adaptive) assessments on a weekly basis. Students were required to achieve 75% of the available points in each mandatory test and were granted three attempts to pass the test within the weekly period.

In order to encourage targeted test preparation, thematically organised practice exercises with progressively increasing difficulty levels were offered. In order to compare the effectiveness of adaptive versus conventional practice formats, we set up three weeks with adaptive exercises (weeks 3,4 and 7) and three weeks with conventional exercises (weeks 5,6 and 9). In week 8 we presented both adaptive and conventional exercises. Each week had 2 to 3 exercise units for separate topics.

Conceptualisation

The **scenario** was adaptive exercise selection for maximally targeted test preparation - each student should work at their level of ability, until their ability estimate reaches the passing threshold for the week's mandatory test.

Based on positive experiences from previous pilot studies, where four to five **difficulty levels** had proved effective, we opted to use four levels in this study, as a finer differentiation was not deemed necessary. Students were advised that they should aim for exercise results above level 3, given that the exam itself would be at level 3 and require 75% of the points to pass.

For **feedback**, students were provided with their estimated level after the test, as well as the opportunity to review all questions together with their own answer and the correct answer (including elaborate teacher feedback on it) after each exercise attempt.

Content and Use

The **question pool** was adapted from the above mention extensive pool of STACK questions for use in foundational mathematics courses. When designing the adaptive exercises, we tested two methods of assigning difficulty: one based on scaffolding and the other based on instructor-assigned difficulty estimates. Scaffolding proved to be significantly less labour-intensive and resulted in more objective difficulty gradation, which is why it was used for the majority of the tasks (see Section 3.2; Fig. 5 shows an item from the Field Test).

We used the **MOUSE plugin** in its optimised version, including question-specific feedback, to create the exercise instances. Each adaptive instance had a question pool of 10 to 30 questions (totalling 171 questions) of which 4 to 6 were administered by the algorithm in each attempt. The instances were used by 20 students who participated in the class activities and took the weekly exams.

Quality Control

Making use of our implemented notification feature, at the end of the term we also reviewed the question difficulty classification. From the total of 171 questions included in the adaptive exercise pools for the three "adaptive weeks", 99 questions were used less than three times, too little for analysis. The remaining 72 questions were reviewed. Among theses, 43 exhibited a rate of correct answers between 25% and 75%, which aligns with the expected trend toward a 50% correctness rate as the algorithm consistently aims to select questions for each participant that they have a 50% chance of solving correctly. Questions that are answered correctly too frequently may be indicative of an inappropriately low difficulty level, while those with very low correctness rates may belong to a higher difficulty category.

This analysis enabled us to identify the questions that deviated significantly from the expected pattern and provided a basis for a targeted review. Among the remaining 29 questions outside the 25–75% range, a vast majority of 23 were answered correctly more than 75% of the time, suggesting they should be reclassified into a lower difficulty tier. This again aligns with Hamamoto Filho et al. (2020), as our question categorisation similarly reflects a tendency toward overestimating question difficulty. Rather than adjusting the assigned difficulty level, we preferred modifying the task slightly to increase its complexity, e.g. by de-scaffolding. The remaining 6 questions with unexpectedly low success rates of less than 25% might have reflected a lack of student understanding. As they all occurred on lower difficulty levels (4 on level 1, 2 on level 2), they were suggested for an instructional intervention in class.

Evaluation

At the end of the class, we asked for qualitative feedback on the adaptive exercises as well as a comparison between the conventional and adaptive exercises. 18 of the 20 students responded. We also present a numerical evaluation of the numbers of exercises taken for each week and success on the mandatory test during that week.

Qualitative feedback The students answered a questionnaire comparing their experience with the adaptive and conventional exercises using a 5-point Likert scale. As shown in Fig. 6, half of the students reported being better able to apply their knowledge and skills in the adaptive tests. This reflects that the algorithm successfully matched questions to individual participants' abilities. At the same time, most students perceived both test types to be equally hard and their results matched their expectations at least as well for the adaptive as for the conventional tests.

Students seemed to have a somewhat better learning experience with the adaptive format: More students found the adaptive test to be more motivating and encouraging

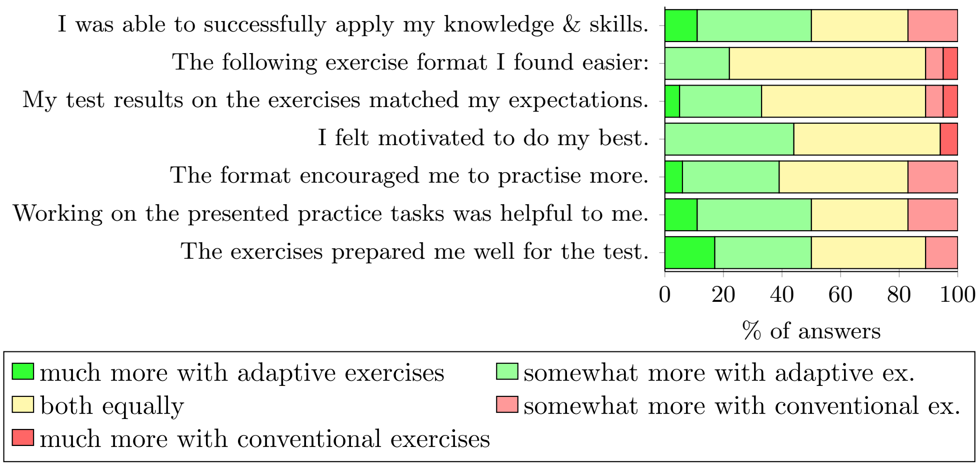


Fig. 6: Student feedback on preferred exercise format. $N = 18$.

them to practise than the conventional tests, although about half the students showed no preference at all. Regarding the learning outcomes, half of the students reported that the adaptive exercises proved more helpful and prepared them better for the weekly exam and only 10% of students felt better prepared by the conventional test.

Fig. 7 shows that half of the students clearly prefer adaptive exercises, while another third indicates no strong opinion, and only few favour conventional tests. This makes a strong argument for the introduction of adaptive tests.

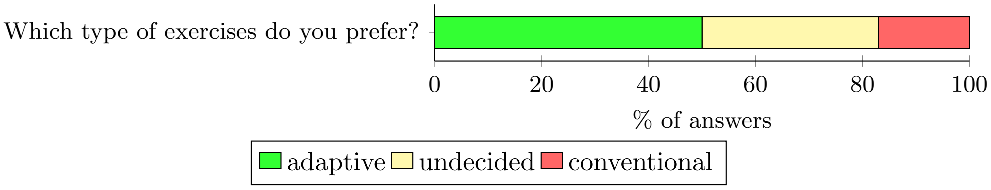


Fig. 7: Student feedback on overall exercise format preference. $N = 18$.

Quantitative results We consider the amount of preparation done in the adaptive and conventional preparation settings, and compare it to the pass rates at first try for the mandatory weekly test. We compare the three adaptive exercise weeks (3, 4, 7 of 14 weeks) to the three immediately neighbouring conventional exercise weeks (5, 6, 9).

We find that in both the conventional and the adaptive settings, about 50% of students used the exercise tests for preparation. They did roughly the same amount of exercise tests on average across the three weeks of each format - 2.1 adaptive exercises per student compared to 2.3 conventional exercises. While the effort taken is approximately the same, students seem better prepared using the adaptive exercise tests: 71% on average pass the mandatory test at first try in the adaptive exercise weeks, as compared to 55% in the conventional exercise weeks. This is apparently not an effect of sequence - the highest number of students passed immediately in week 4 of the class (an adaptive exercise week), followed immediately by the least number in week 5 (a conventional exercise week). We surmise that in the adaptive tests the students are engaging with questions on the correct level of difficulty and are better prepared as a result.

6 Discussion and Conclusions

We have presented our lessons learnt while implementing CAT for exercise selection at the university undergraduate level. We have structured our findings alongside the Will, Skill, Tool, Pedagogy dimensions (Knezek and Christensen 2015), where we want to lower the Skill level needed to adopt adaptive learning for exercise selection through the Tool of the MOUSE Moodle plugin. By applying CAT principles, the MOUSE plugin dynamically adjusts question difficulty based on student responses, ensuring questions are appropriately challenging and thereby supporting students' progression from lower ability levels to greater competency. Our modifications focus on aligning the tool with formative assessment objectives, emphasising iterative learning rather than performance measurement, turning CAT into MOUSE (Module Optimised for Usability in Student Exercises)[6]. Our second contribution towards this goal is the description of the MOUSE process[7], which contains both technical instructions and recommendations on the Pedagogy level for preparing and using adaptive exercises regardless of implementation.

We believe that adaptive exercises can serve student populations with diverse previous knowledge, especially in the beginning of their undergraduate studies. The quality of question difficulty estimates remains crucial, and we support educators in regularly revising the estimates. In order to allow students the informed and self-determined use of the MOUSE plugin, we have stressed that educators should communicate clearly to their students (1) what to expect from the MOUSE tests, (2) what type of feedback to expect and when, and (3) what the desired performance level is and how students' performance relates to it. This allows students to adapt their strategies and use the feedback to optimally guide their individual learning.

Students and educators gave positive feedback on the MOUSE exercises during development and in the MOUSE field test: For example, one instructor stated that the adaptive tests, from a didactic perspective, provide a better match to the students' level and thereby—at least subjectively—increase student motivation, which is also confirmed by interviews with the students. Furthermore, instructors confirmed that the technical implementation works flawlessly and that the time required to create the tasks (at least when using scaffolding) remains within a reasonable range. Further our results suggest (at small N) that the adaptive exercise setup was both qualitatively and quantitatively at least as effective and satisfactory to the students as the conventional tests, while it was clearly preferred by 50% of students. We are currently working with colleagues within and outside our institution to set up MOUSE exercises in their classes. We are also currently working on making the test-end feedback to students more flexible and individual.

[6] https://transfer.hft-stuttgart.de/gitlab/knight/adaptivetests

[7] https://wiki.hft-stuttgart.de/display/ap4/Nutzerleitfaden+Adaptive+Tests

Declarations

- Funding: This work was funded by Bundesministerium für Bildung und Forschung (BMBF), grant 16DHBKI072 (project KNIGHT).
- Competing interests: The authors have no competing interests to declare that are relevant to the content of this article.
- Ethics approval: Not applicable
- Consent to participate: Students were observed during regular course participation; they were free to opt into the study and gave informed consent for their anonymised data to be used in research.
- Consent for publication: Not applicable
- Availability of data and materials: No
- Code availability: The MOUSE plugin is published as Open-Source Software at https://transfer.hft-stuttgart.de/gitlab/knight/adaptivetests
- Authors' contributions: Conceptualisation: UP; Pilot study and field test: AK, KM; Writing: UP, KM, AK

References

Anderson, L.W., Krathwohl, D.A. (eds.): A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's. Pearson Education, Harlow (2014)

Bressoud, D.M., Carlson, M.P., Mesa, V., Rasmussen, C.: The calculus student: insights from the Mathematical Association of America national study. Int. J. Math. Educ. Sci. Technol. **44**(5), 685–698 (2013) https://doi.org/10.1080/0020739X.2013.798874

Bruner, J.S.: The role of dialogue in language acquisition. In: Sinclair, A., Jarvelle, R.J., Levelt, W.J.M. (eds.) The Child's Concept of Language. Springer, New York (1978)

Downey, S.: Test development. In: Peterson, P., Baker, E., McGaw, B. (eds.) International Encyclopedia of Education, pp. 159–165. Elsevier, Oxford (2010)

Frey, A., Hartig, J., Moosbrugger, H.: Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests. Diagnostica **55**, 20–28 (2009) https://doi.org/10.1026/0012-1924.55.1.20

Fink, A., Spoden, C., Frey, A., Naumann, P.: Kriteriumsorientiertes Adaptives Testen mit der KAT-HS-App. Diagnostica **67**(2), 110–114 (2021) https://doi.org/10.1026/0012-1924/a000268

Hamamoto Filho, P.T., Li, E., Ribeiro, Z.M.T., Hafner, M.L.M.B., Cecílio Fernandes, D., Bicudo, A.M.: Relationships between Bloom's taxonomy, judges' estimation of item difficulty and psychometric properties of items from a progress test: a prospective observational study. São Paulo Med. J. **138**, 33–39 (2020) https://doi.org/10.1590/1516-3180.2019.0459.R1.19112019

Harrison, C., Loe, B.S., Lis, P., Sidey-Gibbons, C.: Maximizing the potential of patient-reported assessments by using the open-source concerto platform with computerized adaptive testing and machine learning. J. Med. Internet Res. **22**(10), 20950 (2020) https://doi.org/10.2196/20950

Knezek, G., Christensen, R.: The Will, Skill, Tool model of technology integration: Adding Pedagogy as a new model construct. In: Demetrios G. Sampson, D.I. J. Michael Spector, Isaías, P. (eds.) Proceedings of the 12th International Conference on Cognition and Exploratory Learning in Digital Age (CELDA), pp. 84–91 (2015)

Kibble, J., Johnson, T.: Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? Adv. Physiol. Educ. **35**, 396–401 (2011)

Kim, M.-K., Patel, R., Uchizono, J., Beck, L.: Incorporation of Bloom's taxonomy into multiple-choice examination questions for a pharmacotherapeutics course. Am. J. Pharm. Educ. **76**(6) (2012)

Lord, F.M.: Applications of Item Response Theory to Practical Testing Problems, 1st edn. Lawrence Erlbaum Associates, Hillsdale (1980)

Lilley, M., Pyper, A., Wernick, P.: Attitudes to and usage of CAT in assessment in higher education. Innov. Teach. Learn. Inf. Comput. Sci. **10**, 28–37 (2011)

McCallum, S., Milner, M.M.: The effectiveness of formative assessment: student views and staff reflections. Assess. Eval. High. Educ. **46**(1), 1–16 (2021) https://doi.org/10.1080/02602938.2020.1754761

Oppl, S., Reisinger, F., Eckmaier, A.: A flexible online platform for computerized adaptive testing. Int. J. Educ. Technol. High. Educ. **14**(2) (2017) https://doi.org/10.1186/s41239-017-0039-0

Padó, U., Knebusch, A., Mehmedovski, K.: Computer-based methods for adaptive teaching and learning. In: Uriel R. Cukierman, E.V. Michael Auer (ed.) Advanced Technologies and the University of the Future, pp. 297–317. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-71530-3_19

Raina, V., Gales, M.: Question Difficulty Ranking for Multiple-Choice Reading Comprehension. arXiv (2024). https://doi.org/10.48550/arXiv.2404.10704

Russell, S., Norvig, P.: Artificial Intelligence – A Modern Approach, 3rd edn. Pearson Education, Harlow (2016)

Tiemeier, A., Stacy, Z., Burke, J.: Using multiple choice questions written at various Bloom's taxonomy levels to evaluate student performance across a therapeutics sequence. Innov. Pharm. **2**(2) (2011)

Ueno, M., Miyazawa, Y.: IRT-based adaptive hints to scaffold learning in programming. IEEE Trans. Learn. Technol. **11**, 415–428 (2018) https://doi.org/10.1109/TLT.2017.2741960

van der Linden, W.: Item response theory. In: Peterson, P., Baker, E., McGaw, B. (eds.) International Encyclopedia of Education, pp. 81–88. Elsevier, Oxford (2010)

Wright, B.D.: Practical adaptive testing CAT algorithm. Rasch Measurement Transactions **2**(2) (1988)

Yang, A.C.M., Flanagan, B., Ogata, H.: Adaptive formative assessment system based on computerized adaptive testing and the learning memory cycle for personalized learning. Comput. Educ.: Artif. Intell. **3**, 100104 (2022) https://doi.org/10.1016/j.caeai.2022.100104