Usability design and evaluation for a formative assessment feedback

Florian Horn[1], Daniel Schiffner[2], Thorsten Gattinger[3], Patrick Sacher[4]

Abstract: In a current research project we implemented an approach for delivering computer assisted adaptive testing as an additional form of formative assessment for a lecture. We primarily used these tests as a formative assessment during a "fundamentals of computer programming" course. The tests also included an individual feedback to further guide the students. To evaluate and improve the formative assessments, we performed a usability study, which was focused on the user satisfaction while taking a test and reading the feedback. The study used the user experience questionnaire. The results indicate that an adaptive assessment can provide more support, but also shows shortcomings of the current implementation.

Keywords: formative assessment, adaptive testing, UX, UEQ

1 Introduction

The rising numbers of students at German universities makes a strongly digitalised lecture style and a focus on self-guided learning a necessity. A part of these requirements can be approached by implementing a computer-assisted adaptive test (CAT) as a formal assessment. A CAT is a test that adapts itself to the users' performance, delivering harder questions to users who perform well, as well as easier questions to users who perform badly.

As research shows, an elaborate feedback is crucial for students trying to estimate their learning progress (see [WB96] and [Wa08]). Furthermore, it has to be ensured that the students' user experience is not a hindrance to their learning process.

In this work we evaluate an assessment used in a beginners lecture for programming. We explain our methodology, present our results that were gathered in a comparison study.

[1] Goethe University, studiumdigitale, Robert-Mayer Str.10, 60325 Frankfurt, horn@studiumdigitale.uni-frankfurt.de

[2] DIPF | Leibniz Institute for Research and Information in Education, IZB, Rostocker Straße 6, 60323 Frankfurt, schiffner@dipf.de

[3] Goethe University, studiumdigitale, Robert-Mayer Str.10, 60325 Frankfurt, gattinger@studiumdigitale.uni-frankfurt.de

[4] Goethe University, studiumdigitale, Robert-Mayer Str.10, 60325 Frankfurt, sacher@studiumdigitale.uni-frankfurt.de

2 Related Work

There are different kinds of formative assessments. The survey [MY17] illustrates several, typically used types: Simple multiple choice quizzes, one-minute papers, ePortfolios, student response systems and many other web 2.0 tools.

Furthermore, the survey mentions several implemented formative assessment systems and their usage. Klecker used the learning management system (LMS) Blackboard in 2003 to deliver online multiple choice quizzes to students. These quizzes were instantly checked for correctness and feedback was given immediately [Kl07]. A different approach is taken by CompAssess. CompAssess is a tool for customized assessments with Microsoft Office programs and was evaluated by Brink and Lautenbach [BL11]. It was positively reviewed but had some technical flaws. With PsyCAL (Psychology Computer-Assisted Learning) Buchanan developed a system for multiple choice questions with an instant feedback for wrong choices [Bu00]. The feedback contains excerpts from textbooks to help students understand their wrong answers. For post-graduate students in engineering Burrow et al reviewed three different systems for online formative assessments. Of these, TRIADS (Tripartite Interactive Assessment Delivery Tool) was the best-evaluated system [Bu05]. It may contain up to 40 different interactive questions and provides the correct solutions, as well as the required information, as feedback. The systems evaluation determined it to not be user friendly, but superior to several of its competitors. We used a similar approach of providing students with literature recommendations to revise after the assessment. Rather than providing text excerpts we provided citations.

Other approaches to formative assessments include curriculum-based measurements like the platform quop [Sc19a]. The target audience are pupils from first to sixth grade and the assessed skills are reading, English language and mathematics, starting from fifth grade. After several successful pilot projects this platform was used by over 2000 teachers and 30000 pupils in Hesse [Sc19a]. Lai et al did a user experience study of a web-based formative assessment system for English proficiency [LCC17]. The participants were 28 college students. Their work showed that, given several feedback options in a learning environment (immediate feedback for single questions, delayed feedback for a group questions or immediate feedback for several questions), the option with the highest usability rating was to give a collected feedback for a set of questions. We also used this approach in our formative assessment. The User Experience Questionnaire [LHS08], a questionnaire consisting of 26 Likert scale questions designed to measure the subjective user experience of a software product. The questions consist of opposites and users are tasked with deciding which is more fitting for the software, e.g.: whether the software is more “predictable” or “unpredictable”. These 26 questions are then grouped into 6 categories, each summarising a key aspect of the user experience: attractiveness, perspicuity, efficiency, dependability, stimulation and novelty. For each of these categories the means as well as the standard deviations were calculated and the results compared to the UEQ Benchmark data [SHT17], which consists of 401 studies from a broad range of software products. We performed our usability study using the UEQ.

3 Implementation

Our formative assessment is implemented as a web based training and distributed to students as a SCORM module [SC19b]. Each assessment consists of 15 questions selected from a question pool. The questions were calibrated using the item response theory [Lo80]. The adaptive algorithm targets for a difficulty that allows a student to answer half of the questions correctly.

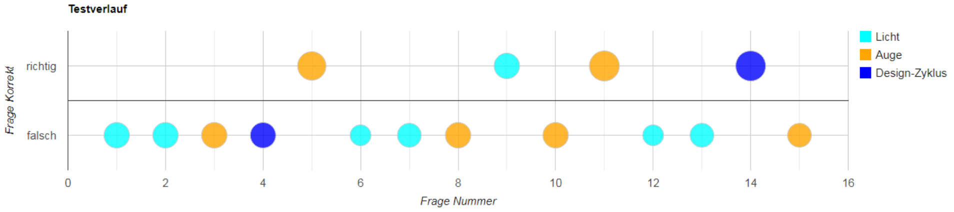

Fig. 1: The visualisation for used in feedback type B. It encodes which questions the student answered correctly (top or bottom), which subject it belongs to (colour) and the difficulty (size, the larger the easier)

Each assessment closes with a feedback presenting an overview of the performance. The basic version (type A) contains only textual feedback, including several performance indicators, a likelihood to solve other questions and literature recommendations. These recommendations varied with the students performance. A struggling student was given basic literature for the lecture, while a competent student received literature for further reading. The feedback does not include results of individual questions. We restrained from giving feedback for individual questions due to a conflict of interest in our design. Originally, we planned to use the assessment both formatively, as well as summatively. Thus, we decided not to leak solutions to questions directly.

We further enhanced the feedback using an additional visualisation (type B), which displays the test run. Correct and incorrect questions are displayed separately, the subject of the question is displayed and the relative difficulty of the question is encoded as well, as explained in figure 1.

We choose to evaluate the effect of including a visualisation in the feedback, since our assessment tool has a basic design and the first informal feedback of students and colleagues indicated that the visual stimulation was low. In addition we were asked for more granular feedback about single questions.

4 Evaluation

We conducted a two-fold usability study of our assessment. First, we questioned our students using short questionnaires and open feedback. Secondly, we performed a comparison study of the feedback types with different students. The latter study utilised the user experience questionnaire (UEQ) to get quantitative data. We split our assessment and its feedback to ensure that the evaluation of our feedback prototypes is independent.

Each session was planned for 15 minutes. Students were presented with a short version of a formal assessment, consisting only of 4 base questions types. Afterwards, the first UEQ had to be filled. The objective of this UEQ was to assess the general performance of the assessment tool, without the feedback, and the results are given in 3. Then they are randomly sent to one of feedback versions. The feedback is shown based on a simulated test. Under the premise that it was their feedback, the second UEQ has been filled.

5 Results

The result of the open evaluation (N=569) was largely considered a net positive. Students enjoyed being kept in flow, making the quizzes equally challenging while still improving their knowledge.

There were two points of critique. The major one was the feedback, especially the inability to tell the correctness of the answers. Students consider it a hindrance when using the formative assessment to study. The literature recommendations were criticised as well, as they were to unspecific.

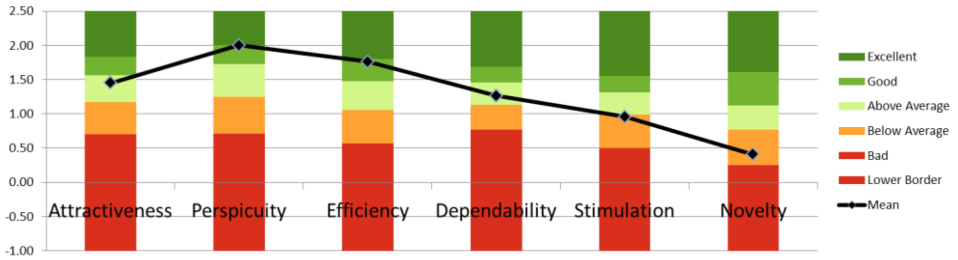

Fig. 2: UEQ Benchmark Results. The mean line illustrates how our software, i.e. the formal assessment tool, performed compared to the 401 existing benchmarks in the UEQ Benchmark, illustrated via the bar colours [SHT17]. N=85

The usability study (N=85) was evaluated using tools supplied with the UEQ [Us20]. As seen in figure 2, we performed excellently in perspicuity, good in efficiency, above average in attractiveness and dependability and below average in stimulation and novelty.

Our main goal was to design a simple and easy to use assessment. This is reflected by our performance in perspicuity, efficiency, attractiveness and dependability. As our design is very basic, our performance in stimulation and novelty is low.

A comparison of the feedback types is illustrated in figure 3 using the UEQs major scales.

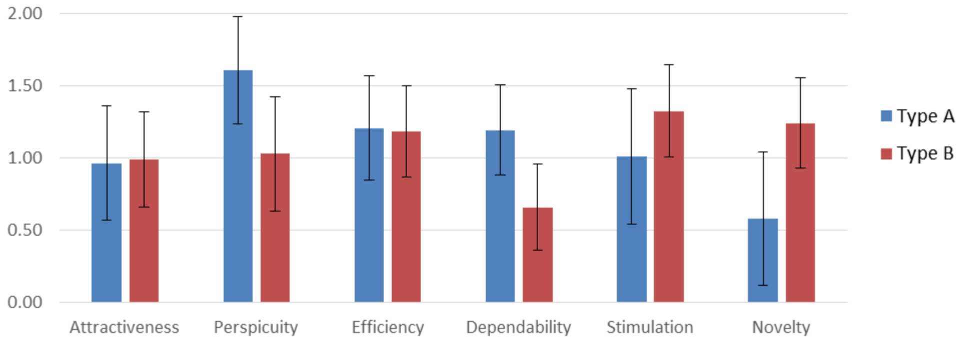

Fig. 3: Results of the A-B comparison study, Type A is the feedback iteration that has been shipped to programming students. Type B is the feedback containing information about single correct answers and the visualisation. N=85, 43 participants received Type A, 42 received Type B.

A Student’s T-Test, for alpha = 0.05, showed that the changes in perspicuity, dependability and novelty were significant. The changes in perspicuity and dependability were unexpected and contrary to the design goal we had in mind. Also, students are rarely confronted with bubblecharts, which improves the novelty but also decreases the readability of the visualisation. Similar results can be found in [BBG18]. We therefore plan to simplify the visualization.

6 Conclusion and Future Work

Our results indicate that the user experience of a formal assessment and its feedback are equally important and should be evaluated separately.

While being considered helpful, we also discovered several challenges when creating feedback. Our visualisation decreased the overall user satisfaction but increased the novelty.

We plan to enhance the feedback with a layered visualisation to allow for simpler representations while still providing more insights. In addition we plan to only use our assessment as a formative assessment in the future and thus will be able to give further guidance, concerning questions and answers, to students.

Bibliography

[BBG18] Bull, S.; Brusilovsky, P.; Guerra, J.: Which Learning Visualisations to Offer Students?: 13th European Conference on Technology Enhanced Learning, EC-TEL 2018, Leeds, UK, September 3-5, 2018, Proceedings. pp. 524–530, 01 2018.

[BL11] Brink, R.; Lautenbach, G.: Electronic assessment in higher education. Educational Studies, 37(5):503–512, 2011.

[Bu00] Buchanan, T.: The efficacy of a World-Wide Web mediated formative assessment. Journal of Computer Assisted Learning, 16(3):193–200, 2000.

[Bu05] Burrow, M.; Evdorides, H.; Hallam, B.; Freer-Hewish, R.: Developing formative assessments for postgraduate students in engineering. European Journal of Engineering Education, 30(2):255–263, 2005.

[Kl07] Klecker, B. M.: The impact of formative feedback on student learning in an online classroom. Journal of Instructional Psychology, 34(3):3, 2007.

[LCC17] Lai, T.; Chen, P.; Chou, C.: A user experience study of a webbased formative assessment system. In: 2017 International Conference on Applied System Innovation (ICASI). IEEE, pp. 899–902, 2017.

[LHS08] Laugwitz, B.; Held, T.; Schrepp, M.: Construction and evaluation of a user experience questionnaire. In: Symposium of the Austrian HCI and Usability Engineering Group. Springer, pp. 63–76, 2008.

[Lo80] Lord, F. M.: Applications of item response theory to practical testing problems. Routledge, 1980.

[MY17] McLaughlin, T.; Yan, Z.: Diverse delivery methods and strong psychological benefits: A review of online formative assessment. Journal of Computer Assisted Learning, 33(6):562–574, 2017.

[Sc19a] Schulen in Hessen : quop. https://www.quop.de/de/quop-in-der-praxis/ schulen-in-hessen, 2019. Accessed: 15/03/2019.

[SC19b] SCORM. https://scorm.com, 2019. Accessed: 15/03/2019.

[SHT17] Schrepp, M.; Hinderks, A.; Thomaschewski, J.: Construction of a Benchmark for the User Experience Questionnaire (UEQ). IJIMAI, 4(4):40–44, 2017.

[Us20] Usability Experience Questionaire. https://www.ueq-online.org/, 2020. Accessed: 18/03/2020.

[Wa08] Wang, T.: Web-based quiz-game-like formative assessment: Development and evaluation. Computers & Education, 51(3):1247–1263, 2008.

[WB96] Wiliam, D.; Black, P.: Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? British educational research journal, 22(5):537–548, 1996.