Progress Testing Anytime and Anywhere – Does a Mobile-Learning Approach Enhance the Utility of a Large-Scale Formative Assessment Tool?

<table>
<thead>
<tr>
<th>Journal</th>
<th>Medical Teacher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>CMTE-2019-1642.R1</td>
</tr>
<tr>
<td>Manuscript Categories</td>
<td>Articles</td>
</tr>
<tr>
<td>Date Submitted by the Author</td>
<td>31-May-2020</td>
</tr>
<tr>
<td>Complete List of Authors</td>
<td>Karay, Yassin; University of Cologne, Faculty of Medicine Reiss, Birger; University of Cologne, Faculty of Medicine Schauber, Stefan; University of Oslo Faculty of Medicine, Centre for Health Sciences Education</td>
</tr>
<tr>
<td>Keywords</td>
<td>Progress testing &lt; Assessment, Computer-based &lt; Assessment, e-learning/computers &lt; Teaching &amp; Learning, Independent &lt; Teaching &amp; Learning</td>
</tr>
</tbody>
</table>
Title

Progress Testing Anytime and Anywhere – Does a Mobile-Learning Approach Enhance the Utility of a Large-Scale Formative Assessment Tool?

Authors

Yassin Karay, Dean’s Office for Student Affairs, Faculty of Medicine, University of Cologne, Cologne, Germany

Birger Reiss, Dean’s Office for Student Affairs, Faculty of Medicine, University of Cologne, Cologne, Germany

Stefan K. Schauber, Centre for Health Sciences Education, Faculty of Medicine, University of Oslo, Oslo, Norway

Name, address, telephone and fax number, and email address of corresponding author

Dr. Yassin Karay
Dean’s Office for Student Affairs, Faculty of Medicine, University of Cologne
Josef-Stelzmann-Str. 20, Geb. 42
50931 Köln (Germany)
Tel.: 0049 – 221 – 478 89217
Fax: 0049 – 221 – 478 88786
Email: yassin.karay@uk-koeln.de
Practice Points

- We provide evidence that the acceptability among students is increased by a location and time flexible version of the formative progress test.

- Test scores were largely comparable across both groups, while time-on-test was different.

- The examination condition had an effect on student’s response behaviour.

- The mobile version of the formative progress test might be related to an increased willingness of many students to take time for the test and to create the best conditions for themselves.
Abstract

Background: The widespread use of mobile devices among students favours the use of mobile learning scenarios at universities. In this study, we explore whether a time- and location-independent variant of a formative progress test has an impact on the students’ acceptance, its validity and reliability and if there is a difference in response processes between the two exam conditions.

Methods: Students were randomly assigned to two groups of which one took the test free of local or temporal fixations, while the other group took the test at the local testing centre under usual examination conditions. Beside the generated test data, such as test score, time-on-test and semester status, students also evaluated the settings.

Results: While there was no significant effect on the test score between the two groups, students in the mobile group spent more time on the test and were more likely to use the help of books or online resources. The results of the evaluation show that the acceptability among students is increased by a mobile version of the formative progress test.

Conclusions: The results suggest that the acceptance and motivation to participate in formative tests is enhanced by lifting local and temporal restrictions. The mobile version nonetheless does not have an impact on the students’ performance.
Introduction

With the establishment of mobile devices such as smartphones or tablets, time- and location-independent learning (in short: mobile learning) has for years increasingly moved into the focus of all areas of the education system (Zeng and Luyegu 2011). The higher education sector is particularly well suited for mobile learning options, as portable devices are widely used among students (Cheon et al. 2012). Market research predicts that 80 to 90 percent of people in Western Europe will own a smartphone by 2018 (Zenith Mobile Advertising Forecast 2017). In Germany, for instance, the dissemination rate in age groups relevant for universities is 95 to 97 percent (Bitkom Research Statista 2018). The Corona Pandemic 2020 has also made a significant contribution to promoting time- and location-independent learning at universities. Many universities worldwide had to offer a digital semester for their students within a very short time. Therefore, it can be assumed that the dissemination rate among students could be increased once again and is almost 100 percent.

The term mobile learning has not yet been defined uniformly. On the one hand, mobile learning is regarded as a part of e-learning (Traxler 2005). On the other hand, mobile learning is seen as an independent learning method (Park et al. 2012). Importantly, a distinct feature of mobile learning is that learners themselves can decide when and where they learn (Peters 2007). Hence, the underlying technology is an enabling factor for offering a more flexible – and cost effective – use of learning resources. Disadvantages might include technical limitations, such as screens that are too small, or slow network speed (Cheon et al. 2012). Further, several authors mention the distraction of the environment or social media applications as a disadvantage (Gikas and Grant 2013). On the other hand, studies show that mobile learning solutions can efficiently support learners in acquiring new knowledge and skills (for an overview see Klimová 2018 and Pimmer et al. 2016). It can therefore make
sense for university lecturers to rethink conventional learning processes and, if necessary, supplement or even replace them with mobile learning approaches.

Indeed, mobile learning approaches can provide educational institutions the opportunity to broaden the instructional toolkit. For instance, some lecturers routinely use mobile devices to administer quizzes or question students' understanding in a lecture or seminar. In this way, mobile learning platforms can augment the existing instructional approach. Mobile learning approaches might also offer a new, more large-scale way to enhance the utility of specific assessment tools and their role in the educational environment. For instance, mobile solutions might play a critical role when medical schools aim at a tight alignment of instruction, student learning and assessment. The use of a mobile learning device in the context of formative testing might be of particular interest. In the particular context of medical education, a type of assessment that is attracting increasing interest and might benefit form a mobile learning approach is progress testing.

Progress testing has been used as a form of both summative and formative longitudinal assessment since it was first developed in the US in the 1970s (Arnold and Willoughby 1990). A key element of progress testing is its longitudinal character. Hence, a progress test assesses an individual’s development of knowledge across his or her course of study. In the Netherlands, Maastricht University was the first to deal intensively with the concept of progress testing (van der Vleuten et al. 1996). Since then, an increasing number of reports mention the use of progress tests in medical schools internationally. In brief, progress testing aims at mitigating usually negative effects of summative assessments (cramming, focusing on isolated topics) by including content from the entire curriculum in each test. Importantly, how often students have to sit a progress test varies between two and four times per academic year.
One reason for the wide adoption of progress testing may be that medical schools can use such a test to pursue several objectives (Wrigley et al. 2012). Indeed, progress tests tend to maximize the benefits of standardized tests by providing reliable and valid information on students’ knowledge, while, at the same time, giving information on the educational environment, useful for curricular planning and improvement. On the downside, however, progress tests are costly: conducting progress tests and maximizing beneficial outcomes is only possible when medical schools invest substantial resources.

In a formative, low-stakes scenario, a major challenge for the beneficial use of progress testing is motivating students to take serious interest and to put effort into the progress test (Schuettelpelz-Brauns et al. 2018). Both students and faculty need to invest in taking the test to receive meaningful, valid feedback from it. Hence, one main objective in research on the formative use of progress testing has been to identify the factors that help or hinder students’ acceptance of the format. Furthermore, a number of studies have aimed at providing empirical evidence for the validity of formative progress tests.

Previous research has found empirical support for the claim that test scores from formative uses of progress tests provide a meaningful source of information. Karay and Schuber find substantial relations between scores on formative progress tests and grades from high-stakes national licensing examinations, implying that a formative progress test might be used as a prognostic/forecasting instrument for the performance of cohorts (Karay and Schuber 2018). Further, an earlier study indicated that direct feedback on test results can increase the acceptance of a formative progress test among students (Karay et al. 2012).

In this study, we are interested in the effects of using a mobile learning environment to provide a formative progress test and hence enabling students to take the test at any time and any location. We conducted an experimental study and compared a mobile to a ‘standardized’
mode of test delivery with respect to a number of critical outcomes. Importantly, van der
Vleuten’s Utility framework (van der Vleuten 1996) informed our research. In this
framework, an assessment’s utility can be understood as a combination of the reliability,
validity, educational impact, acceptability, and the cost effectiveness. More recent
developments, as discussed in (Norcini 2011, Norcini 2018), use a similar perspective to
define what a “good assessment” is. Using this framework as a starting point, we address the
following research questions:

1.) Does the use of a mobile-testing platform increase student acceptability as opposed to
the testing-centre scenario?

2.) Is there a difference in response processes between mobile and standardized testing
scenarios?

3.) Do the different testing scenarios affect the progress test’s generalizability and
reliability?

These questions relate to aforementioned domains of acceptability, validity, and reliability.
We address the two remaining factors - educational impact and cost effectiveness - in the
discussion of this study’s general findings.

Methods

Educational Context

The German-language formative progress test is a knowledge test for undergraduate
medical students designed at Charité – Universitätsmedizin Berlin. Several German-speaking
medical faculties participate in the Berlin Progress Test (BPT). The BPT consists of 200
interdisciplinary MC questions in single-best-answer format. The test covers content from
clinical and para-clinical subjects, as well as from the basic sciences. Every question aims at
testing knowledge relevant for a newly certified doctor (Nouns and Georg 2010). As is typical
for progress testing, on each occasion all participating students across all semesters answer
the same set of items, but items are exchanged between testing occasions. In order to account
for the fact that more junior students are not able to answer a large proportion of the
questions, there is a “don’t know” option that can be used by the test taker. A correct answer
is counted as one point while an incorrect answer is marked negatively (deduction of one
point); the ‘don’t know’ option does not lead to a point deduction. The test score is then
calculated by subtracting the incorrect from the correct answers. Hence, the possible test score
ranges from a minimum of -200 to a maximum of 200.

The practical integration of the BPT into the existing curriculum varies among
cooperating institutions. In Cologne, the BPT is compulsory from the first semester on. In the
four-semester pre-clinical study section the BPT must be taken at least three times and in the
six-semester clinical study section at least five times (for medical curriculum in Cologne see
Zims et al. 2019). Furthermore, the BPT has been carried out computer-based since winter
semester 2009/10. At the end of the test, students receive immediate feedback on their test
results and are able to compare their answers with the right ones. At the end of the semester,
students receive an individual score report prepared by the Charité Medical School. The score
report includes information on personal knowledge development in comparison to one's own
cohort and detailed feedback on one's own strengths and weaknesses in the different
disciplines.

Participants and Procedure
A total of N=1,701 students from semester 2 to semester 10 were randomly assigned to two groups. The first group - the ‘mobile’ group - was assigned to take their progress test using a mobile learning approach. Students in the mobile group could take the test at any place and any time they deemed appropriate. Furthermore, they were asked to do the test on their own, but there were no controls of whether or not they did adhere to this instruction. Importantly, the only limitations in the mobile group were that students had to complete the test in one single session and within a limited time frame (first twelve weeks of the new semester). In contrast, the second group of students – the testing center group– took the progress test as usual. That is, they were invited on specific dates to the local computerized testing center where they took the test under the same standardized conditions as applied to regular high-stakes exams (i.e., no use of books or other materials, no access to the internet, no use of smartphones etc.).

In both conditions, the test was identical with regard to the set of questions used. Furthermore, the progress test and the online questionnaire were delivered using ILIAS software (i.e., an open source learning management system; https://www.ilias.de/en/about-ilias/) which could be accessed using a standard internet web browser (mobile group) or a secure exam browser (learning centre group). This ensured equal testing platforms, independent of the specific hardware used.

At the end of the semester, about two weeks after the mobile test, the two groups were asked separately about their subjective attitude to their version of the BPT. Students were invited via e-mail to fill out an anonymous online questionnaire that used both comments and scales to rate their experiences.

Measures
Test performance and test-taking behavior. Two objective measures of test-taking behavior were included in this study. First, we used number-correct scores on the progress test as well as the so-called test-score, that is, the number of correct answers minus the number of incorrect answers. Typically, both scores are highly correlated and can usually be used interchangeably for the research objectives of this study. Second, we recorded the time a particular student sat the test, in minutes. Based on these measures, students are usually classified as ‘unmotivated’ and ‘motivated’ test-takers. This criterion is a standard approach in the Berlin Progress Test. Participation is described as ‘unmotivated’ if a student hands in the test in less than 20 minutes or chooses the don’t-know option on all 200 questions. Details on this procedure are given in Nouns and Georg 2010 and Nouns et al. 2012.

Self-reported motivation and perceived utility. We administered a self-report questionnaire in order to triangulate students’ perceptions with the more objective data (score, time). Importantly, in alignment with the local data protection regulations, students were not obliged to answer these forms; data is obtained anonymously and individual responses are not linked to the participants' actual test results. A total number of N=9 questions focused on issues of test-taking behaviour, acceptability and testing conditions. All questions used a rating scale in five categories ranging from “I completely agree” to “I do not agree at all”. The open questions were counted with respect to the number of positive and negative comments. We chose a number of typical statements given by the students to illustrate what we deemed to represent a positive or a negative comment.

Analysis

For testing the first research question, that is, whether the use of mobile-learning increased acceptance as opposed to the standardized scenario, we analyze differences across the two groups in regard to the self-report questionnaires, the open responses on these
questionnaires. When considering differences in responses on the administered questionnaires, we use a generalized linear model to test for differences between groups simultaneously. In a typical regression context, such a group effect would be included as a predictor for some kind of outcome variable. It is, indeed, also possible to include group membership as the outcome variable in a logistic regression. The advantage of this approach in our context is that instead of comparing multiple variables independently of each other, they can be compared simultaneously in one regression model; this avoids issues of multiple comparisons that would be evident when independently testing those variables. Furthermore, we analyze between-group differences in the propensity of individual students being classified as “unmotivated” as opposed to “motivated” test takers by applying a generalized linear model.

The second research question focuses on investigating a possible effect of the different testing conditions on students’ response processes. Using a linear mixed model, we address this question by comparing how the time students used on the test (dependent variable) was related to performance (independent variable) and whether there were differences in this respect between the groups. We furthermore test an interaction effect, that is, we account for the possibility that the association between the time used and the performance itself differs between the two groups (performance-group interaction). Using data from the self-reports, we investigate differences between the two groups on variables concerning the response process.

The third question is related to whether the change in the mode of test administration did impact the reliability of the test scores. For this we estimated variance components for the following aspects: students, items, semesters, and their respective interactions as well as the residual variance term. We calculated the relative share of each of the random effects estimated. In addition, we calculated the within-semester reliability coefficients (Cronbach’s Alpha) across the two groups. Taken together, we investigate whether there were systematic
differences in the generalizability of the results that were related to the according testing condition.

Results

Descriptive Statistics

In both groups, there were about 70% female participants. Age was comparable with $M_{(age, \text{mobile})} = 25.3$ years as opposed to $M_{(age, \text{testing centre})} = 25.1$ years. Tables 1a and 1b give more detailed descriptive statistics on the key variables included, both for the full dataset and the dataset including only students classified as ‘motivated’ or ‘serious’ test takers. In general, students scored better the further they were in their studies. Furthermore, across semesters, students sat the test longer when they were in the mobile group.

For the self-reported motivation and perceived utility questionnaires, complete data for $N = 341$ students was available. The majority of responses was given by students in the mobile group, from which data for $N_{\text{mobile}} = 213$ students were available. A number of students commented on the test using open commentaries, but did not respond to the questionnaires. In the mobile group, $N = 78$ students gave an answer using open commentary; $N = 35$ of such comments were obtained from the testing centre group.

Research Question 1: Does the use of a mobile-testing platform increase student acceptance as opposed to the testing-centre scenario?

Differences in evaluation commentaries. As noted above, we tested for differences in evaluation from comments simultaneously by using a generalized linear mixed effects model that included semester as a random effect. Group was entered as the dependent...
variable. The items in the questionnaire were predictor variables. In this way, the approach integrates the comparisons across the variables at once, instead of necessitating multiple tests. Figure 1 provides the predictor variables in increasing order, and detailed results for all administered items are found in Table 2. We found that the more strongly students agreed with the question that they could take the progress test under best possible conditions, the more likely they were to be in the mobile-group (OR = 18.29, CI 7.24 – 46.23, \( p < .001 \)). Furthermore, students who disagreed more strongly with the counter-factual question that the other condition would have been more motivating were more likely to be in the mobile-group (OR = 0.13, CI 0.05 – 0.34, \( p < .001 \)).

The open commentaries were categorized into positive, negative and neutral evaluations of the Progress Test. A total of N=78 free text comments were given in the mobile group: of these 73 were positive comments, 1 was negative and 4 were neutral. By contrast, in the testing centre group, 35 comments were given in total. Of these, 21 were positive, 3 were negative and 11 were neutral. In total, there were more positive (N = 94) comments than negative ones (N = 4). A corresponding chi-square test found a statistically significant difference between the number of type of comments given (\( \chi^2 = 155.55, \text{df} = 2, p < 0.001 \)). There was, however, no statistically significant difference between the mobile and the testing centre group concerning this pattern of type of comments (\( \chi^2 = 5.80, \text{df} = 2, p = 0.05 \)).

**Differences in likelihood of being classified as a “non-serious” test taker.** The results from a generalized mixed effects model including semester as a random effect indicated that the likelihood of ‘non-motivated’ participation (i.e., only marking the ‘don’t know’ option; handing in the test in a short time while not scoring sufficiently) was 2 times higher in the testing center group than the mobile group (OR = 2.02, CI 1.59 - 2.52, \( t = 5.71, p < .001 \)).
We excluded participants who were classified as ‘non-motivated’ from further analyses. Hence, in the following analyses, \( N_{\text{mobile}} = 544 \) students from the ‘mobile group’ were included as compared to \( N_{\text{testing center}} = 367 \) in the ‘testing center’ condition (see Tables 1a and 1b).

**Research Question 2: Is There a Difference in Response Processes Between Mobile and Standardized Testing Scenarios?**

**Differences in test-scores across groups.** A linear regression model was used to estimate the group effect, the semester and their interaction. We found an effect for semester (\( \beta = 10.00, \ CI \ 9.03-10.96, \ t=20.34, \ p<.001 \)), that is, scores increased as a function of how advanced students were in their studies (standardized \( \beta = 0.66 \)). Furthermore, neither the group effect (\( \beta = -3.81, \ CI \ -13.18 – 5.56, \ t = -0.80, \ p = .426 \)) nor the group-semester interaction (\( \beta = -0.12, \ CI \ -1.63 – 1.38, \ t = -0.16, \ p = .874 \)) were statistically significant. We cross-validated the results using a linear mixed effects model and accounting for the different group sizes by including a random effect for semester. Estimates were only marginally affected. In summary, the findings indicate that students that are more senior scored, on average, higher than more junior students (cf. Figure 2). We did not find evidence favoring the interpretation that this relationship could be affected by the mode of administration (‘mobile’ vs. ‘testing center’).

**Time-on-test differences between groups.** Again, a linear model was used to estimate fixed effects for the group effect, the semester and their interaction. In this case, we found a statistically significant effect for both semester (\( \beta = 3.81, \ CI \ 2.82 – 4.81, \ t = 7.48; \ p < .001 \)) and group (\( \beta = -21.16, \ CI \ -30.88 – -11.44; \ t = -4.27; \ p < 0.001 \)). Put differently, the more advanced students were in their studies, the longer they sat the test. In addition, students
in the mobile group sat the test longer than students in the testing centre group. The
interaction was not statistically significant ($\beta = -0.68$; CI -2.24 – 0.89; $t = -0.85$; $p = 0.397$).

**Differences in test-taking behaviour.** In the next step, we analysed the data with
respect to whether there was a difference between groups in how time-on-test and
performance (i.e., correct answers on the progress test) were related to each other. This was
done by using a generalized linear mixed model, including semester as a random effect. The
number of correct answers was entered as the dependent variable and time, group, and their
interaction as fixed effects. The results (Table 3) indicate that time used was positively related
to number of correct answers across groups ($\beta=0.71$, CI 0.59 – 0.82, $t=12.16$, $p<.001$). In
addition, students in the testing-centre condition answered more questions per minute
correctly as compared to students in the mobile condition, as indicated by a significant time-
group-interaction ($\beta=-0.22$, -0.35 – -0.09, $t=-3.43$, $p =.001$). Finally, there was a non-
significant main effect for the group ($\beta = 4.89$, CI -2.91 – 12.68, $t = 1.23$, $p = 0.219$). The
results of this analysis are presented in Figure 3, which illustrates the interaction effect (i.e.,
non-parallel regression slopes).

Results from the analysis of students’ responses to the self-report questionnaire align
with the findings from the regression model. Students in the mobile group were more likely to
agree to the statement that they used help in the form of books or online resources for
answering questions in the BPT (OR = 4.02, CI 1.61 – 10.03, $p = 0.003$).

**Research Question 3: Is There a Difference in Score Reliability?**

Again, we used a mixed effects model to estimate variance components. In this
analysis, we used data at the level of responses to the particular questions. We included
random effects for Students, Items, Semester and Group in the model. The largest variance component was the residual variance, accounting for 61.20% of the total variance. Students, Items, Semester, and Group accounted for 9.90%, 14.30%, 9.60%, and 0.10%, of the total variance, respectively. The only notable interaction was that of Semester and Items, accounting for 4.60% of the total variance. Detailed results are given in Table 4.

Cronbach’s alpha was $\alpha=0.98$ in the mobile group and $\alpha=0.98$ in the testing centre group, across all participating students. Within semesters, Cronbach’s alpha varied between 0.93 and 0.98 in the mobile group and between 0.91 and 0.97 in the testing centre group (Table 5). In summary, the analysis of variance components did not indicate any systematic differences in the group effect. Importantly, estimates of score reliability were largely comparable across the two testing scenarios.

**Discussion**

In this study, we investigate whether a time- and location-independent variant of a formative progress test would affect the test’s utility using an experimental design. In particular, we focus on questions of acceptability (RQ1), validity (RQ2) and reliability (RQ3). Our results indicate that the acceptability among students is increased by a location- and time-flexible version of BPT. In the evaluation, both the mobile group and the testing centre group clearly favour the mobile version. The majority of students in the mobile group who responded to the evaluation forms stated that they were able to create the best conditions for processing the BPT by choosing their own time and location. Probably because of the more flexible way to handle the BPT, more students participated in the mobile condition than in the testing centre group.

The possibility to complete the BPT regardless of time and place might be related to an increased willingness of many students to give themselves time to take the test. This might
point at a difference in response processes between the groups. Indeed, compared to the
testing centre group, the mobile group reported that they took more time to answer the
questions in the best possible way. This statement corresponds to the increased actual time
used on the test and the analysis regarding motivated participation. Students from the mobile
group worked longer on the test than the other group. Furthermore, they were less likely to
quit taking the BPT than students from the testing centre group.

Our analyses indicate that the percentage scores were comparable across groups while
time-on-test was different. In a regression analysis, we found that the testing centre students
correctly answered more questions per minute than the students in the mobile group. This
finding might be explained by the fact that, as students indicated in the self-report
questionnaires, they were more likely to use resources to answer questions. Taken together,
these findings suggest that the condition had an effect on student’s response behaviour. At the
same time, scores were largely comparable across both groups.

In view of the increasing competition among universities for financial resources, the
consideration of the economic and effective use of scarce resources also plays an important
role in teaching. Universities that want to succeed in the face of intensified competition must
increase their performance, efficiency and flexibility. Van der Vleutens Utility framework
also considers cost efficiency. Compared to the conventional variant of administering the
progress test in a testing-centre under exam conditions, the mobile variant of the BPT is
significantly more cost-effective. For the mobile variant, for example, there is no need for
personnel to supervise the test. Neither is it necessary to provide the rather expensive
infrastructure (rooms and computers). In our context, 18 appointments of 3 hours each are
typically required to enable all students in the 2nd to 10th semesters to take the test. Due to
the limited number of computers in the testing centre, a total number of 18 days of testing had
to be organized and per day, two to three staff members supervised the test.
Evidently, shifting to a mobile platform of assessment might change how students and institutions use both the test itself and the test scores. For instance, students in the mobile group might have used the progress test as an opportunity to learn rather than as an actual test or exam. Indeed, in the mobile group, almost a quarter stated that they had sought help when answering the questions on the test. Still, students have a choice of how to take the test and how to work through the questions, which may be a key factor in using the test in a way that enhances learning. Furthermore, research in cognitive psychology and learning science has long shown that the context in which learning occurs is particularly important for learning (Godden & Baddeley, 1975; Anderson, Reder & Simon, 1996; Frankland, Josselyn & Köhler, 2019). One way of enhancing learning is, according to this strand of research, to diversify the context of learning and assessment. In this way too, students might benefit from not taking the exam under standardized testing conditions.

While we indeed speculate that learning could be enhanced by diversifying assessment contexts, another function of the BPT might be impaired. Since the test is no longer taken under standardized exam conditions, using test scores for benchmarking across institutions is more problematic. However, comparisons are generally limited in this scenario, since the BPT’s integration into instructional settings varies across institutions. Therefore, differences and similarities in results can only be compared and discussed against the background of possible influencing factors.

A limitation of the present work is certainly that our study is limited to a single institution. In addition, the mobile testing approach delineated in this study might affect the test’s validity and educational impact beyond the effects described. Furthermore, the self-report questionnaire was composed of statements directly tied to the specific testing scenario here. Its construction was not based on any published or validated framework, but rather motivated by how the test is embedded in the specific context. Hence, we assume that this set
of questions displays a sufficient degree of face validity. Importantly, the general findings from both the analyses of objective measures and the self-report are rather congruent. For this reason, we argue that the two sources complement each other and that self-reports critically add the students’ perspectives on the intervention discussed here.

Further limitations are related to the deploying tests and other instructional tools on mobile devices, in general. Critics of mobile learning often cite environmental distraction as a disadvantage. In the context of the current study, this point of criticism could not be confirmed from the perspective of students. The students in the mobile group stated that they were less distracted by the environment (e.g. by telephone, other people, etc.) than students in the testing centre group. Students of the mobile-group described that they can create a much quieter working environment for themselves. Some free text comments of the students give information about the reasons. For example, some students suggested that concentration is not permanently interrupted by fellow students who, for instance, prematurely leave the examination room due to non-serious test-taking.

For the progress test considered here, as much as for other low-stakes testing endeavours, a major challenge is to motivate students to use the test as a valuable feedback tool. Mobile technologies might play a crucial role in offering more flexibility and thus more autonomy in times of crowded schedules and exam dates. Indeed, this might be among the reasons that mobile learning in general is quite popular with medical students. In addition, against the background of increasingly heterogeneous lifestyles of students, universities have to consider the diversity of their students (Karay et al. 2018). In our example this means that commuters, students abroad, working students, and students with children can use the mobile variant to independently determine the optimal time and place for taking the test. Put differently, it is likely not just the ‘mobile’ nature of the intervention itself that affects the tests overall utility. Rather, this approach introduces some beneficial ‘underlying’ aspects...
such as an increase in students’ autonomy and flexibility of where, when and how (i.e., using resources) to take the test. Hence, further research is needed to isolate the specific beneficial aspects of the mobile learning approach. Nevertheless, with the help of flexible teaching and assessment concepts, it is possible to better integrate the students’ environment and the educational context - and hence, we argue, to foster learning.
References


Karay Y, Schauber SK. 2018: A validity argument for progress testing: Examining the relation between growth trajectories obtained by progress tests and national licensing examinations using a latent growth curve approach, Medical Teacher, DOI: 10.1080/0142159X.2018.1472370.


### TABLES

Table 1a: Descriptive statistics for all participants

<table>
<thead>
<tr>
<th>Semester</th>
<th>Mobile Group</th>
<th>Testing Centre Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Testscore (correct-incorrect)</td>
</tr>
<tr>
<td>2</td>
<td>88</td>
<td>11.69</td>
</tr>
<tr>
<td>3</td>
<td>94</td>
<td>9.04</td>
</tr>
<tr>
<td>4</td>
<td>58</td>
<td>10.57</td>
</tr>
<tr>
<td>5</td>
<td>86</td>
<td>27.55</td>
</tr>
<tr>
<td>6</td>
<td>78</td>
<td>25.81</td>
</tr>
<tr>
<td>7</td>
<td>94</td>
<td>39.53</td>
</tr>
<tr>
<td>8</td>
<td>78</td>
<td>43.31</td>
</tr>
<tr>
<td>9</td>
<td>87</td>
<td>51.36</td>
</tr>
<tr>
<td>10</td>
<td>56</td>
<td>47.54</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>719</td>
</tr>
</tbody>
</table>

Table 1b: Descriptive statistics for participants classified as ‘serious’ test takers

<table>
<thead>
<tr>
<th>Semester</th>
<th>Mobile Group</th>
<th>Testing Centre Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Testscore (correct-incorrect)</td>
</tr>
<tr>
<td>2</td>
<td>73</td>
<td>13.59</td>
</tr>
<tr>
<td>3</td>
<td>74</td>
<td>12.7</td>
</tr>
<tr>
<td>4</td>
<td>42</td>
<td>16.93</td>
</tr>
<tr>
<td>5</td>
<td>76</td>
<td>31.36</td>
</tr>
<tr>
<td>6</td>
<td>57</td>
<td>36.46</td>
</tr>
<tr>
<td>7</td>
<td>69</td>
<td>52.86</td>
</tr>
<tr>
<td>8</td>
<td>55</td>
<td>61.45</td>
</tr>
<tr>
<td>9</td>
<td>60</td>
<td>72.92</td>
</tr>
<tr>
<td>10</td>
<td>38</td>
<td>69.21</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>544</td>
</tr>
</tbody>
</table>
Table 2: Generalized linear mixed effects model for the responses to evaluation forms

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Odds Ratios</th>
<th>CI</th>
<th>Statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.22</td>
<td>0.11 – 0.45</td>
<td>-4.17</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>&quot;I took sufficient time for answering the questions&quot;</td>
<td>1.09</td>
<td>0.57 – 2.10</td>
<td>0.27</td>
<td>0.789</td>
</tr>
<tr>
<td>&quot;I used available resources (books/internet) for answering the questions&quot;</td>
<td>4.02</td>
<td>1.61 – 10.03</td>
<td>2.98</td>
<td>0.003</td>
</tr>
<tr>
<td>&quot;I would like to use old tests for training purposes&quot;</td>
<td>2.10</td>
<td>1.09 – 4.03</td>
<td>2.23</td>
<td>0.026</td>
</tr>
<tr>
<td>&quot;I got interrupted while taking the test (e.g., by others, telephone calls)&quot;</td>
<td>2.34</td>
<td>1.29 – 4.22</td>
<td>2.81</td>
<td>0.005</td>
</tr>
<tr>
<td>&quot;The progress test should be done self-administered and on a mobile platform&quot;</td>
<td>1.53</td>
<td>0.86 – 2.69</td>
<td>1.46</td>
<td>0.145</td>
</tr>
<tr>
<td>&quot;...in addition it should be possible to take the test under regular exam conditions&quot;</td>
<td>0.77</td>
<td>0.43 – 1.40</td>
<td>-0.85</td>
<td>0.393</td>
</tr>
<tr>
<td>&quot;The mobile (alternatively: testing center) condition provided the best opportunity for creating optimal conditions for working through the test&quot;</td>
<td>18.29</td>
<td>7.24 – 46.23</td>
<td>6.14</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>&quot;As a formative test, the progress test should be used for rehearsing high-stakes testing and thus be done under summative exam conditions&quot;</td>
<td>0.50</td>
<td>0.31 – 0.82</td>
<td>-2.74</td>
<td>0.006</td>
</tr>
<tr>
<td>&quot;It would have been more motivating to do the test on a mobile platform (in the testing center)&quot;</td>
<td>0.13</td>
<td>0.05 – 0.34</td>
<td>-4.19</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Random Effects

| σ² | 3.29 |
| T₀₀ Fachsemester | 0.00 |
| Observations | 341 |

Note. Questions 7 and 9 were differently phrased for the two groups.
Table 3: Linear mixed effects model for test score on the formative progress test as a function of assigned group and time on the test

<table>
<thead>
<tr>
<th>Predictors</th>
<th>β</th>
<th>β_hat</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>16.29</td>
<td>2.02</td>
<td>0.043</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>0.71</td>
<td>0.60</td>
<td>12.16</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Group</td>
<td>4.89</td>
<td>0.06</td>
<td>1.23</td>
<td>0.219</td>
</tr>
<tr>
<td>Time*Group Interaction</td>
<td>-0.22</td>
<td>-0.26</td>
<td>-3.43</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Random Effects

<table>
<thead>
<tr>
<th>Variance Component</th>
<th>Observations</th>
<th>Estimate</th>
<th>Relative Share (in Percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>911</td>
<td>1.96</td>
<td>9.90</td>
</tr>
<tr>
<td>Group</td>
<td>2</td>
<td>0.02</td>
<td>0.10</td>
</tr>
<tr>
<td>Item</td>
<td>200</td>
<td>2.84</td>
<td>14.30</td>
</tr>
<tr>
<td>Semester</td>
<td>9</td>
<td>1.89</td>
<td>9.60</td>
</tr>
<tr>
<td>Group*Item</td>
<td>400</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Group*Semester</td>
<td>18</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Semester*Item</td>
<td>1800</td>
<td>0.91</td>
<td>4.60</td>
</tr>
<tr>
<td>Semester<em>Item</em>Group</td>
<td>3600</td>
<td>0.05</td>
<td>0.20</td>
</tr>
<tr>
<td>Residual</td>
<td>12.11</td>
<td>61.20</td>
<td></td>
</tr>
<tr>
<td>Semester</td>
<td>Mobile Group</td>
<td>Testing Centre</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>--------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.95</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.93</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.97</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.96</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.97</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.97</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.98</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.98</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.98</td>
<td>0.97</td>
<td></td>
</tr>
</tbody>
</table>
FIGURES

Figure 1. Results from the evaluation questionnaire showing in statements to be endorsed as a function of the condition students were assigned to.

Figure 2. Number of expected correct answers for the two groups across semesters showing generally slightly higher scores for the mobile learning condition (the group effect is not statistically significant)
Figure 3. Time on test and its relation to the number of expected correct answers for the two groups.