Implementing assessment innovations in higher education

Boevé, Anna Jannetje

Chapter

7

Discussion

7.1 Introdcution

As discussed in chapter 1 this dissertation was driven by challenges and dilemmas faced in classroom assessment in higher education. Because the research was conducted in real settings with different possibilities and limitations for research designs, the collection of studies in this thesis were based on different theoretical lenses and used different methodologies. Whenever, due to the nature of the problem, fully experimental designs (RCTs) are not feasible, employing a combination of research methodologies to answer a set of related research questions is advocated (e.g., Carey & Stiles, 2016).

7.2 Summary of the main findings

By means of a field experiment, in chapter 2 I demonstrated that students' performance on exams consisting of multiple choice questions did not differ depending on the mode of examination. The computer-based exam facilities at the time warranted an experimental design in a real-life context, allowing strong conclusions. Given the important role of technology in society, it was somewhat surprising that results of a survey among both a Dutch and an international cohort of students did not find support for computer-based exams; about half of the students still preferred a paper-based exam. Initial results of a qualitative enquiry into the reasons why students preferred a certain mode of examining showed that universities can take measures to reduce the stress students experience in such a high-stakes situation. For example, students prefer to change the layout of the exam according to their own preference (one question at a time, or all questions at once), and they like to have note-taking and other editing functionalities in computer-based exams.

In Chapter 3 I investigated whether the use of subscores on an exam can add additional information to the total scores on an exam. This is a relevant question in exam practice because there is a vivid discussion to what extent assessments can be used to guide or steer students' further learning based on identifying students' strengths and weaknesses. In assessment practice at universities, information about subscores on an exam could, for example, direct students towards parts of the content that they have not mastered so that those who failed can use this information to study for a resit exam. Based on this discussion in practice as well as recent literature (e.g., Sinharay, 2010; Sinharay, Puhan, & Haberman, 2011) the focus in chapter 3 was on the reliability of subtest scores. Although providing students with feedback about their performance based on subscores seems to be a good idea, subtest scores may not be as reliable as the total score of an exam. This was demonstrated both using an exam testing different types of knowledge and an exam consisting of both open - and closed questions. Interesting was that some of the open questions contributed to a better measurement precision, especially the open questions that used more structured scoring (and are thus most straightforward to grade) contributed least to the measurement precision.

The research presented in chapters 4 and 5 involved the use of online technology to improve the learning process. In chapter 4 the implementation of practice tests was explored in different contexts. In the first part, the relationship was investigated between students' use of practice test resources and exam results in two statistics courses and a biopsychology course. Results showed that there was a positive relationship between the

voluntary use of practice tests and performance on the final exam, but the strength of the relationship varied in the different courses. In the second part of this study the performance of cohorts with practice test resources was compared to a cohort without practice test resources by means of test equating. The mean performance of the cohorts with voluntary practice test resources was not higher than the mean performance of the cohort that did not have access to voluntary practice tests.

In Chapter 5 I explored students' study behaviour in a flipped classroom course and a regular course by means of bi-weekly diaries. The aim of implementing the flipped classroom was that students would study more, and more intensely, throughout the course, due to required preparation and active engagement during the lecture. Results from the diaries showed that students' study behaviour in the flipped classroom course was not very different from students in a regular course. Furthermore, study behaviour did not appear to be strongly related to student performance in both the flipped and regular course. Exploration of students' study behaviour in the course evaluations showed that some students experienced the flipped course design as intended to support their learning process. Other students however, demonstrated resistance to changing their study behaviour even though changing study behaviour is expected in order to benefit from the flipped classroom.

Chapter 6 was inspired by the literature on higher education and research conducted in chapters 4 and 5. Researchers in higher education often cannot use randomized controlled trials to evaluate innovations, and, therefore, often use quasi-experimental designs to compare the results when similar courses are followed, or to compare the results of the same course across different years. Student performance in this type of research is compared without taking into account the natural variation in exam scores. Based on all the grades from first year courses over a period of six years, I studied the variation in mean course grades. Overall, about 17% of the variation in grades could be attributed to the year and course level, while almost 40% of the variation in passing a course could be attributed to the year and course level. Using this information I illustrated that a statistically significant difference in course grades may fall within the expected variation in grades. Thus, observing a moderate increase in student performance after the introduction of an educational intervention may not be indicative of the intervention's effectiveness.

7.3 Limitations

The research in this thesis was conducted at a single university in the Netherlands. Therefore, an important limitation was that the results found in the various chapters may not generalize to other programs or higher education institutes, such as other types of higher education (e.g., Higher Professional Education in the Netherlands), or institutes in other countries. However, the issues addressed in the present study do reflect challenges many institutes, in different countries, face.

When implementing an innovation in education, it is often not possible to use an experimental design and, therefore, it is often difficult to evaluate the causal effect of an implementation. Perhaps to the frustration of practitioners, much research, including that presented in chapters 4 and 5, is by design unable to suggest that implementing a flipped

classroom or practice tests lead to better student performance. Perhaps field experiments need to be employed on a much larger scale in order to evaluate the impact of innovations in higher education. This could be achieved in a large well-coordinated project where randomization at the course level is possible. For example, one or more institutions may be involved, with instructors willing to implement an innovation, such as the flipped classroom. A risk in field experiments is that of spillover-effects, that is, effects of the implementation trickling into the participants assigned not to receive the implementation, so that such a design would need to ensure that only courses from a single program of study would be included. Such a large-scale design would be the most convincing, and potentially most generalizable. A sufficiently large sample size at the course level would also be imperative in order to be able to analyse the results adequately (Hox, 2010).

7.4 Scientific contributions

The aim of this dissertation was to empirically evaluate several assessment innovations in higher education. In doing so, the various chapters contributed in different ways to both methodological and theoretical debates in higher education research. With respect to the methodology, in chapter 6 it was discussed that many research projects in higher education innovations focus on the quasi-experimental comparisons of courses that are similar in content, or use subsequent cohorts of students following the same course. If the grade variation of many higher education institutes was known, then perhaps it would be possible to determine how large the beneficial effect of an educational measure should be in order to be larger than a regular variation to be expected purely based on random year-to-year and course-to-course variation. This could greatly enhance the discussions about what kind of effect size is reasonable to work towards and is needed in terms of the expected benefits of the educational innovation. It is reasonable to expect that in order to make causal claims such as “this educational intervention has the desired effect”, more stringent demands should be met compared to correlational claims such as “the mean pass grade after the innovation was statistically significantly higher than the mean pass rate before the innovation”. Other methodological contributions could be found in chapter 3 and in chapter 5, where we used test-equating procedures and recent developments in test theory about subtest scores in classroom assessment. There is not much research that applied these methods in classroom assessments and as we showed they may contribute to better quality research in higher education.

From a theoretical perspective, the testing effect demonstrated by cognitive psychologists (e.g., Roediger & Karpicke, 2006), as well as assessment theory focusing on the feedback functionality (e.g., Hattie & Timperley, 2007) suggest that the learning process can be improved and that better student performance may be expected when practice tests are implemented. Chapter 4, however, showed that it is difficult to study whether this is actually the case, and that in practice student performance did not necessarily improve. This raises an important question for researchers in assessment: if the benefit of formative assessment is not visible or measurable in terms of student performance, how then should its success be evaluated and in what way should practitioners try to transfer the results of scientific research to their teaching practice?

Research in medical education has suggested that the implementation of cumulative or progressive testing may be beneficial (Kerdijk, Tio, Mulder, & Cohen-Schotanus, 2013; Kerdijk, Cohen-Schotanus, Mulder, Muntinghe, & Tio, 2015; Saint, Horton, Yool, & Elliot, 2015). In cumulative testing, students receive multiple exams throughout a given period in such a way that in each test, all the previously taught material is tested (including the material taught before the previous test). This ensures that students continue to study all the relevant material regularly, with each test counting towards the final grade. It is important to note, however, that this form of testing increases the summative function of a test (rather than remaining purely formative), which does reduce some of the benefits formative assessment can have on the learning process. Furthermore, Kerdijk et al (2015) did not find overall improved student performance in a group of students with cumulative tests, compared to a group of students with a final exam only. In addition, it is important to take the broader curriculum into account, as the research on cumulative assessment in medical education (particularly Kerdijk et al., 2013 and Kerdijk et al., 2015) is in the context of an integrated curriculum such that in each study period there are no separate courses. This in contrast to the curriculum of the social and behavioural sciences faculty in the present study, where students follow multiple separate courses at the same time, of which each is assessed separately. More research is necessary to determine whether cumulative testing is an ideal combination of formative and summative assessment functions to improve student learning in higher education.

As another example, several theories such as, active-learning (Prince, 2004), and student-engagement (Ashwin and McVitty 2015; Kahu 2013) all provide a convincing case for why flipping the classroom could be a good idea. Most research, however, attempted to demonstrate improved student performance (e.g., Davies, Dean and Ball 2013; Mason, Shuman and Cook 2013; McLaughlin et al. 2013; Pierce and Fox 2012; Street, Gilliland, McNeill and Royal 2015) rather than the intended behaviour change targeted by implementing the flipped classroom. By focusing on the targeted behaviour change it became evident in chapter 5 that students' self-regulation is also important to take into consideration. The importance and interaction of different theories in educational research will be much better understood when focusing on the targeted change of an educational innovation, rather than on the indirect outcome such as improved student performance. Thus chapter 5 provided an important contribution to the next steps in researching the implementation of the flipped classroom.

7.5 Contribution to practice

Based on the findings in chapter 2, an important practical finding was that the mode of exam administration did not influence the performance on exams consisting of multiple choice questions. However, when transition to computer-based exams also implies changing the type of question (e.g., using constructed response questions instead of multiple choice questions as in Dermo, 2009 and Peterson & Reider, 2002), care should be taken that the assessment give students a fair chance to demonstrate their mastery of the learning goals at the intended levels of knowledge. Given that students do not seem to prefer computer-based exams over paper-based exams, higher education institutes should carefully consider

the design and affordability of different computer-based exam applications. An application demonstrated by McNulty et al. (2007), for example, showed how test-taking strategies students were accustomed to in a paper-based exam could also be applied in computer-based exams. This could increase students control over the test-taking process and help reduce the stress students have during exams

The research conducted in chapter 4 also has several practical implications. When lecturers use incentives such as bonus points or extra credit so that all students benefit from formative assessment, this may have the unfortunate consequence that students do not use the assessment in a truly formative manner (as also shown by Kibble, 2007). On the other hand, keeping assessment truly formative, that is, when it is not part of the grade, means that students are required to be more in charge of their own learning process. This may result in students not using the provided formative assessment resources. There is no definitive way to solve this dilemma, and lecturers need to be aware of these issues when deciding how to implement formative assessment. Furthermore, it is important to take the whole course context into account. In the case of the two statistics courses, the course design already activated students with mandatory small-group practical meetings in addition to the large-scale lectures. In these courses the use of practice tests correlated weakly with student performance on the final exam. In the case of a course with voluntary large-scale lectures only the correlation between students' use of practice tests and the final exam score was much stronger.

When innovations are implemented in higher education with the intention to improve the learning process, the intended and potential benefits need to be communicated with students on a regular basis. Although students were informed of the potential benefits of the flipped classroom, the course evaluations showed that some students did not understand or believe in the potential of the flipped classroom to aid their learning and this might have influenced their decision to study as usual. As in chapter 4, the dilemma of whether to require students to participate in activities that are intended to facilitate learning was also evident in the case of the flipped classroom course in chapter 5.

For higher education institutes there is a difficult trade off: if student responsibility for their own learning is desired, then learning activities do not need to be mandatory. You can lead a horse to water, but you cannot make it drink. This is a particularly difficult principle to hold on to in the context of institutional accountability with performance incentives for students who perform well. Institutions need to reflect critically whether they are offering students all the means to be self-regulated learners, or whether the aim of self-regulation is an excuse not to invest in facilities that may support learning.

To conclude, a further discussion should take place between stakeholders concerning the expected output of educational innovations. If an innovation is implemented to improve the learning process of students, what exactly then is the expected outcome of this improvement and for whom this outcome is important? If the aim is to improve the learning process of students, then evidence should be collected on the learning process. A critical discussion should arise when the expected outcome is better student performance: Should a greater percentage of students pass at the first attempt? Should a greater percentage of students pass overall? Should the mean grade of a cohort increase? These

outcome measures could be informative to evaluate the effectiveness, but care should be taken to define a priori how big of an improvement is to be expected, and evidence concerning the mechanism by which this is achieved should also be collected (i.e. study behaviour, use of innovations) should be collected. With the collected evidence, the use of test-equating strategies, and/or taking into account the variation in grades could be fruitful avenues to then evaluate the results. By combining sources of information, and not exclusively focusing on the outcome of student performance, lecturers and universities may gain more insight in the effectiveness of assessment innovations. A continuing collaboration between research and practice is necessary to improve the quality of assessment and learning in higher education.