Implementing assessment innovations in higher education

Boevé, Anna Jannetje

Chapter

4

Implementing Practice Tests in Psychology Education

Note: Chapter 4 has been submitted as

Boevé, A. J., Albers, C. J., Bosker, R. J., Meijer, R. R., & Tendeiro, J. N. *Implementing Practice Tests in Psychology Education.*

4.1 Introduction

Given the massification of higher education there is a continuing search for how to maintain and improve the quality of the teaching-learning process (Hornsby & Osman, 2014). With large classes of, sometimes, hundreds of students, the teacher-student ratio becomes very small, leaving teachers with little time and resources to monitor their students' learning process. Assessment is understood to be a driver of student's learning process (Gibbs, 1999; Schuwirth & Van der Vleuten, 2011). By using web-based technology, teachers can provide large groups of students with assessments for learning and feedback through, for example, practice tests.

There are still many questions, however, as to how to implement practice tests. An important issue is the extent to which the use of practice resources or tests should be completely voluntary or, whether the use of practice resources should be accompanied with performance incentives. It is also unclear at the university were this study took place to what extent implementing practice tests in a cohort of students was associated with better performance compared to a cohort that was not offered practice tests. Therefore, the present study consists of two parts: in the first study we discuss students' use of practice tests and the association between use and student performance in three different courses that differed in the amount of participation incentives. In the second study, we investigated how two cohorts of students that had access to voluntary practice quizzes performed on the final exam in comparison to a cohort that did not have access to voluntary practice quizzes.

4.1.1 Theoretical Background

An important distinction can be made between summative and formative assessment functions (Black & Wiliam, 2003; Wiliam & Black, 1996). The summative function refers to passing a judgment, whereas the formative function of assessment is meant to aid the process of learning. Examples of summative assessments are when a final exam is used to decide whether a student has sufficiently achieved the learning goals, or when a selection test is used to determine whether a candidate is sufficiently skilled to enter a particular program. In contrast, formative assessment can be considered assessment *for* learning rather than assessment *of* learning (Schuwirth & Van Der Vleuten, 2011). In practice, the boundary between formative and summative assessment is not always clear. The primary aim of practice tests for example is to improve student learning, with the focus on the formative function of assessment although there may be a summative component if the assessment counts towards a grade. Any type of official grading or participation incentive increases the stakes of a test and increases the summative function of a test. On the other hand, tests with a primarily summative function can also offer opportunities for learning when students have the opportunity to see how they performed on different parts of the assessment.

There are various theories to explain why student learning can be improved through formative assessment. The function of assessment to aid the learning process has been theorized to consist of feed-up, feedback, and feed forward (Hattie & Timperley, 2007). Feed-up can be considered the learning goal a student is working towards,

feedback is the evaluation of a student's current standing relative to the learning goals, and feed forward involves determining the next steps to be taken in the learning process. According to Hattie and Timperley (2007) the combination of addressing these questions determines the effectiveness of feedback. Furthermore, de Kleijn, Bouwmeester, Ritzen, Raemakers, and Van Rijen (2013) showed that these three aspects were all instrumental in determining the reason why students voluntarily use of formative assessment in medical education.

Research from cognitive psychology suggests that the very act of retrieving information from memory consolidates the ability to remember information, and this is known as the testing effect (Roediger & Karpicke, 2006). The implication of this finding is that any test, whether summative or formative, can potentially benefit a student's learning process if a student actively has to retrieve information from memory. The testing effect has been consistently found in lab studies, and several authors have argued that it generalizes to educational practice (e.g., Roediger & Karpicke, 2006). Studies claiming to find the testing effect in practice, however, often use evidence from practice questions identical to questions on the final exam (e.g., McDaniel, Wildman, & Anderson, 2012). This can be problematic in practice at universities where it is prohibited to administer the same questions in a practice exam and a final exam. Furthermore, Carpenter et al. (2017) showed that for three out of four exams, there was no positive correlation between the number of practice tests completed and the score on the final exam questions that were not identical to or modified versions of practice test questions.

There is empirical research that corroborates the expectation that the use of practice tests is positively associated with student performance. Carrillo-de-la-Peña et al. (2009), for example, showed that students who participated in voluntary practice test in a proctored exam environment midway through the course, performed better on the final exam compared to students who did not participate. Other studies also found that students' use of quizzes was positively associated with student performance (e.g., Angus & Watson, 2009). In attempting to improve participation of students in formative assessment, students are sometimes given incentives for participation, such as the result of formative assessment counting towards the final grade. However, Kibble (2007) showed that when student's score on the quiz counted towards their grade, almost all students completed the quizzes with a perfect score, but did not perform as well on the actual exam. In this case, the relationship between practice test use and performance on the final exam is weakened.

There may also be another explanation for why the relationship between using practice tests and student performance is not strong. Roediger, Agarwal, McDaniel, and McDermott (2011) found that children who did not use practice tests performed better on the summative tests compared to those who did use practice quizzes, suggesting that children who did not use the practice tests were effectively able to determine whether use of the practice tests would benefit their learning process. The research by De Kleijn et al. (2013), also found that an important reason why students in higher education chose not to use practice tests was that they already sufficiently mastered the course material. The advantage then of providing completely voluntary practice tests, is that it

allows students to regulate their own learning process. In contrast, however, there is also research demonstrating that students overestimate the extent to which they comprehend the material, and that they may stop studying too soon (Karpicke, 2009).

Although the implementation of practice tests is motivated by teachers aiming to improve student learning and thus performance, only a randomized controlled trial could establish such a causal relationship. In practice, and in the present study, it is both unfeasible and unethical to randomly assign part of the students to a condition with access to practice tests and assign another group to a condition without access. This means that educational research often relies on quasi-experimental research to evaluate changes to the learning environment.

Three main approaches have been taken to determine whether implementing practice tests is positively associated with students' performance. First, some research has considered the relationship between the score on practice tests and the score on final exams (e.g., Marden, Ulman, Wilson, & Velan, 2013; Kibble, 2007). When implementing completely voluntary practice tests that may be completed any number of times, however, it is unclear to what extent the final recorded score reflects the performance of a test-taking situation, or whether students filled out the answer with information on hand (i.e., without retrieving information from memory). Therefore, using the final score of the practice tests was not considered appropriate in the present study. A second approach is to examine how often or whether students use practice tests and relate this to the final score on the exams (e.g., Carrillo-de-la-Peña et al. 2009). However, it is important to consider that students who use practice tests may simply be students that are more motivated and conscientious learners than students who do not use practice tests. In meta-analyses, the amount of explained variance in college GPA that was due to motivation and study skills was found to be around 10%–15% (Robbins, Lauver, Davis, Langley, & Carlstrom, 2004; Richardson, Abraham & Bond, 2012). A third way practice test implementation has been evaluated is by considering how different cohorts of students perform, in which there is a non-randomized control group in the form of a cohort that did not receive practice tests (e.g., Dobson, 2008). What has often been neglected thus far in cohort comparison studies, however, is the equivalence of exams in subsequent cohorts. In the present study, we considered both the relationship between student use of practice tests and student performance (study 1), as well as the cohort comparison approach, to evaluate the implementation of practice tests at a university in the Netherlands (study 2).

4.2 Study 1

The goal of the first study was to evaluate the effectiveness of implementing practice tests in three courses in a first-year psychology program at a Dutch university. The courses differed in the number and types of practice tests available to the students. In one course, the practice tests were all completely voluntary, while in two other courses, there was a combination of voluntary practice tests and practice tests for which participation could count towards fulfilling the prerequisites for taking the exam, without actually being part of the course grade. The primary research question of interest was: To what extent did students use these practice tests and is the use of practice test resources related to student achievement on the final exam?

4.2.1 Method

Course characteristics

This study was conducted with students from the University of Groningen in the Netherlands, which has two bachelor psychology programs: an international program taught in English and a program taught in Dutch. The use of practice tests was evaluated in three courses that took place in the academic year 2014/2015: a course in biopsychology in the international program with 400 enrolled students, and two statistics courses, called Statistics 1a and Statistics 1b, in the Dutch program with 330 enrolled students in Statistics 1a, and 333 enrolled students in Statistics 1b. In total 265 students participated in both statistics courses. Some students took only one statistics course as a result of re-taking a course after failing it in the previous year(s).

Biopsychology. The course Biopsychology covered 15 chapters of material spread over a period of seven teaching weeks, with two lectures each week. There were two exams, a midterm exam and a final exam, each consisting of 40 multiple choice questions. The midterm exam covered the first eight chapters, and the final exam covered the final seven chapters of the course material. Students' grades were determined by the combined score on the two exams. Lecture attendance was not mandatory, and there were no required activities or assignments for this course such as practical meetings or homework assignments.

Two types of practice tests were made available as digital tests on the online-learning platform of the university known as Nestor, a local version of Blackboard (www.blackboard.com). The first type of practice test offered to students was a quiz. A total of 30 quizzes (two for each chapter) was available, each containing 15 true/false questions. The second type of practice test was a sample exam for both the midterm and the final exam. The sample exams for the midterm and final exam each contained 40 multiple choice questions. Taking both types of practice tests was voluntary and when students completed the practice tests, they received direct feedback on which questions they had answered correctly (see Figure 4.1).

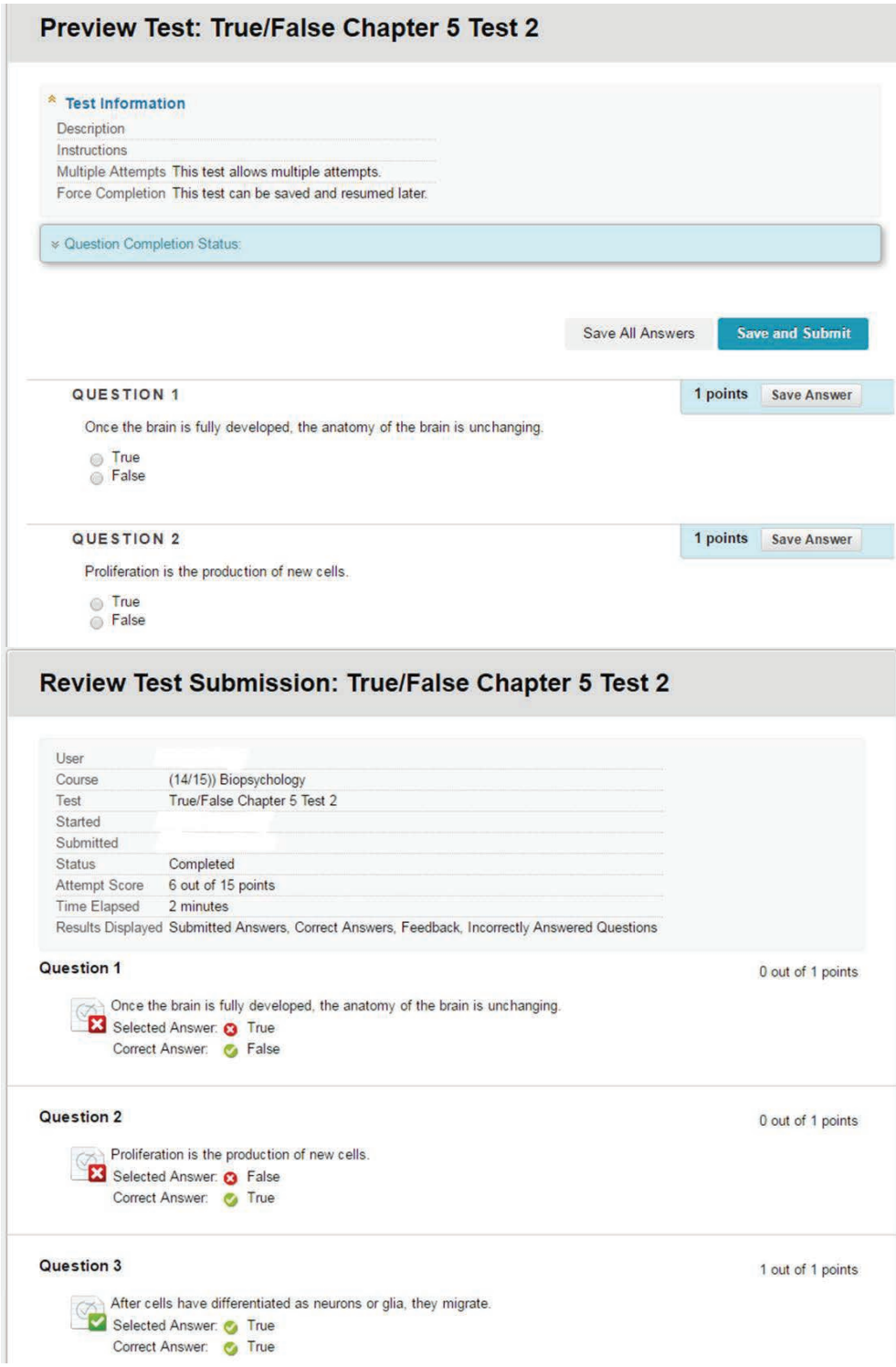


Figure 4.1. Example of practice quiz questions and feedback in Biopsychology.

Statistics 1a and 1b. The courses Statistics 1a and Statistics 1b covered five and four chapters of material from the same textbook, respectively. Both courses had seven teaching weeks, with two lectures each week of which one lecture was used to explain the material, and one lecture was used to answer students' questions. Lecture attendance was not mandatory, as in the biopsychology course. Students were required to attend a weekly practical meeting and hand in homework assignments that were not part of the final grade, but were graded as sufficient or insufficient. In addition, students taking the statistics courses for the first time were required to make use of practice test resources in a portal accompanying the textbook that offered different types of practice tests. The total amount of points possible in the portal was 2,600, and students had to earn at least 1,000 points in order to pass. There were different ways in which students could earn points in the practice portal. Students could choose to make pre-tests and post-tests of the material discussed in the chapters, they could complete tests called "learning curves", or they could answer questions about what happens when manipulating variables in an interactive environment called an "applet". This practice portal was made available and designed by the textbook authors/publisher (courses.bfwpub.com/ips8e). Its use (which practice elements of the portal were made available to students) was tailored to the learning goals of Statistics 1a and 1b by the course instructor.

A combined total of attended practical meetings, a sufficient amount of homework, and a pass for using the practice-test portal was a prerequisite for being allowed to participate in the final exam, which determined students' grades. In addition to the above course design that made use of the practice portal, there were some additional voluntary practice test resources. In Statistics 1a, there was one type of voluntary practice test, which was the sample exam covering all the material of the course and consisted of 30 multiple choice questions. In Statistics 1b, there were two types of practice tests: a sample exam, and four short quizzes for each chapter of the material. The sample exams for both courses, and the quizzes in Statistics 1b were offered through the university's online learning platform as digital tests, in the same way as the biopsychology course. These practice tests were voluntary, and students could complete the quizzes and sample exams when they wanted, and as many times as they liked. Table 4.1 provides an overview of the number and types of practice tests in each course. Students who failed the course in a previous year but did have sufficient practical attendance in that year, were exempted from the practical activities, including use of the practice portal.

Table 4.1 Amount of practice tests offered for each course, by the different types of tests.

	Chapter quizzes	**Sample exam**	**Mandatory practice portal**			
			Pre-test	Post-test	Learning curve	Applet
Biopsychology	30	2 parts	-	-	-	-
Statistics 1a	-	1	5	5	16	3
Statistics 1b	4	1	4	4	12	-

Measures

Student use of practice resources. The practice quizzes for Biopsychology were grouped in a different folder for each lecture week of the course in the online learning environment. Using the online learning environment, it was possible to evaluate the extent to which students accessed the seven folders containing quizzes for each week of the course material, as well as the extent to which students accessed the sample exams. Thus, students use of practice resources for Biopsychology was measured with two variables, one indicating the number of weeks for which quizzes were accessed (0 through 7), and one indicating the number of sample exam parts accessed (0, 1, or 2).

For Statistics 1a and Statistics 1b, there were records of students' scores on the practice quizzes, sample exams, and the number of points earned in the practice portal for each type of completed test (pretests, posttests, learning curves, and applets). Since the voluntary quizzes could be completed an unlimited amount of times and it was not possible to determine how seriously students took the quizzes, the scores on the practice quizzes were not taken into consideration. For Statistics 1a and Statistics 1b, therefore, practice test use was measured by means of whether students had completed each specific practice test (each type in the practice portal, and the voluntary quizzes and sample exams). The total number of each type of practice test completed was calculated and used for the analyses.

Student performance. For all three courses the final exam was a multiple-choice test with 4 answer options for each item. In Biopsychology, both the midterm and final exam consisted of 40 items. In order to receive a passing grade, students had to earn at least 53 out of 80 points. The exams of Statistics 1a and Statistics 1b consisted of 32 and 28 items, respectively. In order to receive a passing grade on the final exam, students in Statistics 1a had to answer 21 out of 32 exam questions correctly, and students in Statistics 1b had to answer at least 19 out of 28 exam questions correctly. For ease of interpretation, the proportion of questions answered correctly by each student was used as the measure of student performance for all three courses.

Analyses

After examining the extent to which students used the practice tests in the different courses, linear multiple regression analysis was conducted to examine the extent to which using practice test resources was related to student performance. For Statistics 1a and 1b, only students who participated in the practice portal were included in the regression analyses. All analyses were conducted in *R* (version 3.4.1; *R* Core Team, 2017). Visual inspection of the residuals gave no reason to doubt the assumption of normality (q-q plot), or homoscedasticity (plots of residuals vs. predictors). The linearity assumption may be somewhat violated for the relationship between student performance and the number of weeks of quizzes accessed by students in the Biopsychology course, leading to a potential underestimation of the amount of explained variance. However, due to the limited amount of observations at certain measurement points we restricted ourselves to the simplest model.

4.2.2 Results

For Biopsychology, out of the 440 students who enrolled, 384 attended both the midterm- and the final exam. For Statistics 1a, 305 of the 330 enrolled students attended the final exam, and for Statistics 1b, 311 students of the 333 enrolled attended the exam. There were 265 students who attended both the Statistics 1a and Statistics 1b exam, and the Pearson correlation between the total scores on both exams of these students was $r = .46$ ($p < .001$).

Practice quizzes were accessed by 90% ($N = 347$) of the students who completed Biopsychology, and 77% ($N = 297$) accessed one or both parts of the sample exam. The sample exam was accessed by 65% ($N = 199$) of students who completed Statistics 1a, and 25% ($N = 78$) of students who completed Statistics 1b. In Statistics 1b, 37% ($N = 115$) accessed the voluntary quizzes. The mandatory practice portal was used by 89% ($N = 272$) of students in Statistics 1a, and 85% ($N = 264$) of the students in Statistics 1b.

Table 4.2 shows that the use of practice resources accounted for about 24% of the variance in student performance for Biopsychology, about 5% of the variance in student performance for Statistics 1a, and 8% of the variance in student performance for Statistics 1b. For the Biopsychology course, both the number of quizzes accessed and completing the sample exam were statistically significant predictors. Table 4.2 shows that for Statistics 1a and 1b, only the use of some practice tests predicted student performance. For Statistics 1a, completing the sample exam and the number of applets accessed were both statistically significant. For Statistics 1b, only the completion of practice quizzes was a significant predictor. Table 4.3 illustrates the difference in mean exam score between students who did not access any, or accessed all of the practice tests that were statistically significant predictors of student performance in the models.

Figures 4.3 through 4.5 provide the distribution of the exam scores as a function of (a) the number of quizzes accessed and sample exams taken for Biopsychology (Figure 4.3); (b) the sample exams taken and number of applets accessed for Statistics 1a (Figure 4.4) and (c) the number of quizzes completed (Figure 4.5). These figures visually confirm the results of Table 4.2, that is, it appears that the more students access quizzes, applets, or spend time in practicing exam questions, the better their results on the exams, although the effect is small. This is especially the case for both statistics courses, where the combined effect of all these variables leads to 3% (Statistics 1a) and 6% (Statistics 1b) explained variance.

Table 4.2 Multiple regression coefficients and model results for student performance predicted by the use of practice tests in each course, with (m) indicating mandatory portal.

	Biopsychology		**Statistics 1a**		**Statistics 1b**	
	B(SE)	*p*	*B(SE)*	*p*	*B(SE)*	*p*
Intercept	66.2 (1.3)		58.2 (2.5)		55.4 (2.5)	
Quizzes	1.8 (0.3)	<.001			2.3 (0.6)	<.001
Sample exam	2.8 (0.9)	.002	3.9 (1.5)	.01	1.7 (2.2)	.43
Pretest (m)			0.1 (0.4)	.81	0.4 (0.7)	.58
Posttest (m)			-0.2 (0.5)	.71	-0.8 (1.0)	.42
Learning curve (m)			-0.1 (0.2)	.66	0.2 (0.2)	.45
Applet (m)			1.7 (0.6)	.02		
R^2/ R^2_{adj}	.24/.23		.05/.03		.08/.06	
***F*(*df*)**	58.90 (2, 381)		2.79 (5, 266)		4.42 (5, 256)	
p	<.001		.018		.001	

Table 4.3 Percentage of students who completed none and all of the different types of practice tests and their mean exam score for the statistically significant predictors of student performance in each course.

		None completed		**All completed**	
	Test-type	*N*(%)	*M*(*SD*)	*N*(%)	*M*(*SD*)
Biopsychology	Quizzes	37 (10)	56.65 (12.62)	179 (47)	68.01 (6.95)
	Sample exam	87 (23)	56.67 (11.52)	197 (51)	66.75 (7.91)
Statistics 1a	Sample exam	106 (35)	18.88 (3.70)	199 (65)	20.15 (3.62)
	Applet	123 (40)	19.23 (3.57)	10 (3)	21.20 (3.61)
Statistics 1b	Quizzes	188 (60)	16.11 (4.06)	47 (15)	19.29 (3.89)

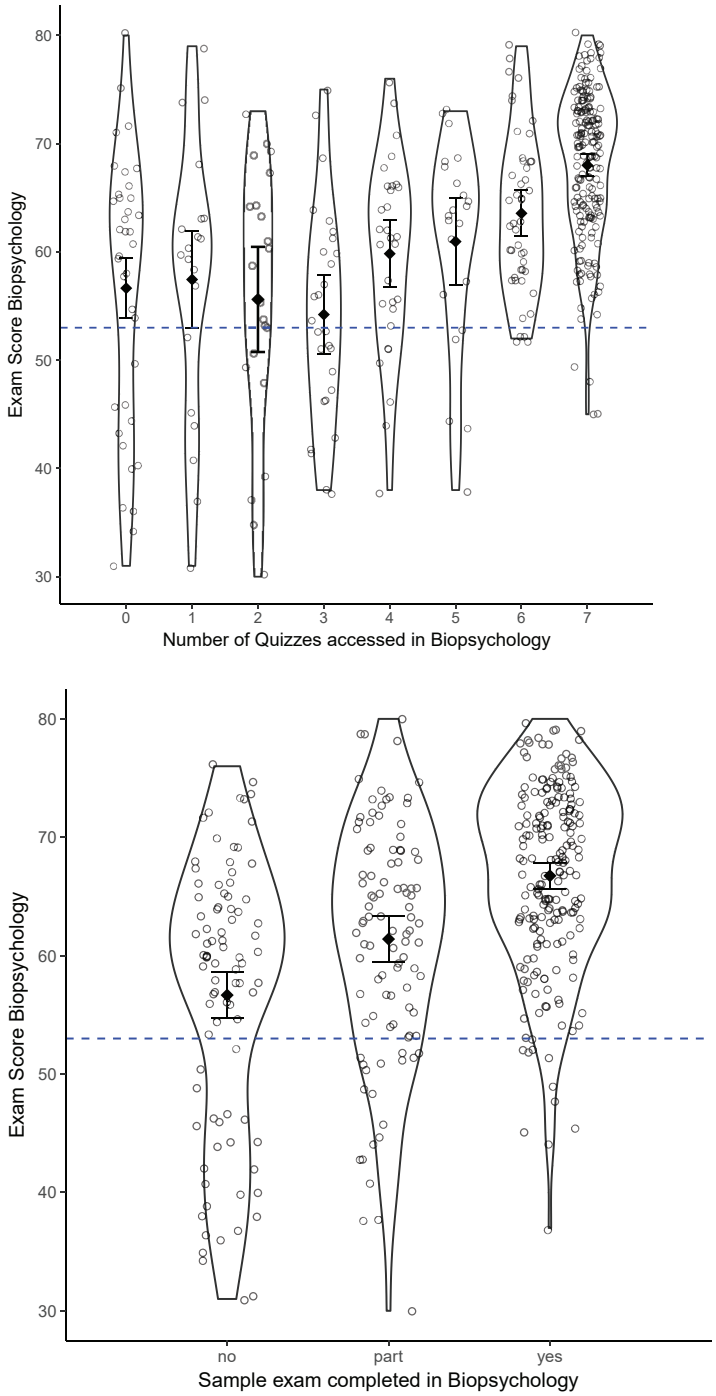
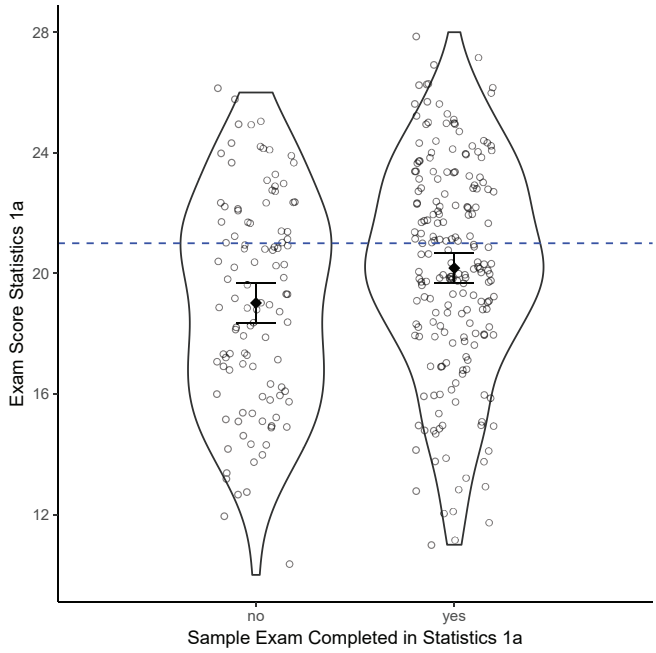


Figure 4.3. Distribution of exam scores by the number of quizzes (left) and sample exam access (right) in Biopsychology, with the dashed line indicating the cut-off score for a pass, and error bar indicating the 95% CI around the mean.



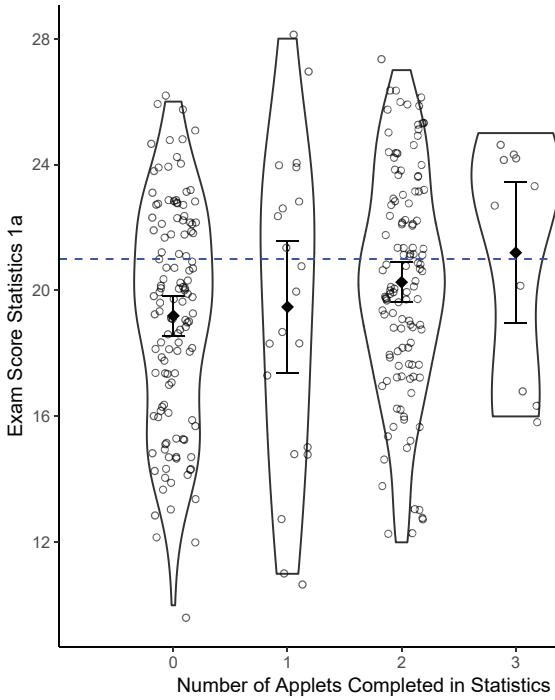


Figure 4.4. Distribution of exams scores in Statistics 1a by whether the sample exam was completed (left) and the number of portal applets (right) accessed with the dashed lines indicating the cut-off score for a pass, and error bar indicating the 95% CI around the mean.

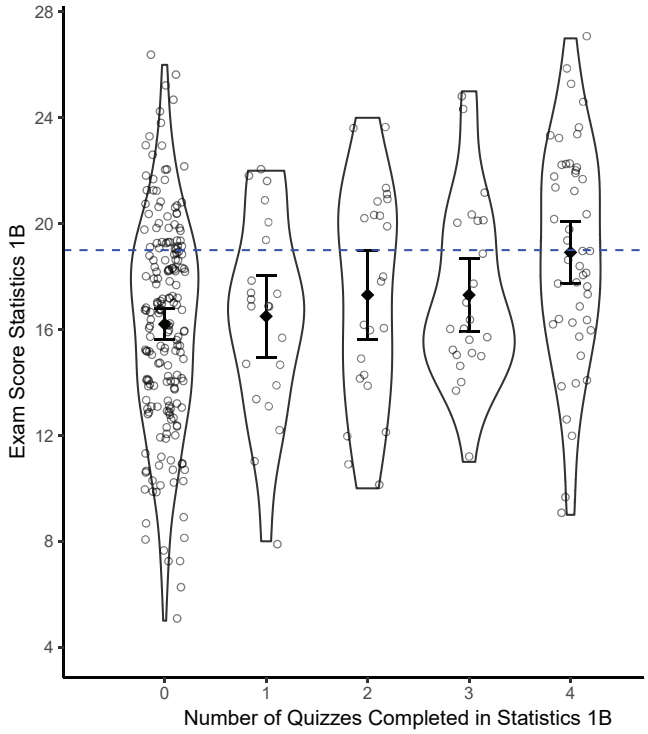


Figure 4.5. Distribution of exam scores by the number of voluntary quizzes completed in Statistics 1b, with the dashed line indicating the cut-off score for a pass, and error bar indicating the 95% CI around the mean.

4.3 Study 2

In the first study, we saw that the relationship between practice test use and student performance was strongest, yet still small, in the Biopsychology course, which also had the highest rate of participation in the practice tests. To gain further insight into whether implementing practice tests is an effective tool to improve student performance, we conducted a follow-up study for the course Biopsychology. In this second study we examined student performance across three different cohorts of which the first cohort did not have access to practice test resources, and the second and third did have access to practice tests. The primary research question of the second study was: do students from cohorts with practice test resources perform better than students from a cohort with no practice test resources? In contrast to prior research, we used item response theory (IRT, see for example, Embretson & Reise, 2000), to equate the exams from the different cohorts so that the scores between cohorts could be meaningfully compared.

4.3.1 Method

Sample. Exam results for the same mandatory course on Biopsychology as discussed in the first study were analyzed from three cohorts of students who were enrolled in the international program of the psychology bachelor program (years 2013, 2014, and 2015). Table 4.4 shows the number of students who completed the final exam for Biopsychology in each year. These numbers include students who may have re-taken the course as a result of failing the course in the past. There were 13 students who completed the exam in both 2013 and 2014, and seven students completed the exam in both 2014 and 2015.

Practice Test Implementation. The courses were designed in a similar way in all three years, except for the implementation of the practice tests. Practice tests were implemented in the international track in 2014 and in 2015 in slightly different ways. In 2014, the practice tests were made available to students via the online learning environment as images (one image for each quiz), with the answer key in a different file. In 2015, the practice tests were made available in the same learning environment in the format of digital tests that students could complete and then received feedback on (see Figure 4.1 as described in study 1). The practice tests were accompanied with an instruction for students on how to use the practice tests. As the aim of the practice tests was to improve the performance of the students, it was expected that both cohorts with practice tests performed better than the control cohort 2013.

Measures. Each exam for all three cohorts originally consisted of 80 multiple choice questions. One item from the 2013 exam was excluded from the analyses since all students answered it correctly. The final number of test items in each exam is shown in Table 4.4. Examination policy required that no two subsequent exams had the exact same questions in order to prevent cheating. Thus, to compare the total scores across the different exams in a meaningful way these scores should be placed on a common scale (that is we should equate the scores). To make this possible the exam of cohort 2015 was designed to include some items from both the exams in cohort 2014 and 2013. This inclusion of so-called anchor items enabled the scores to be equated using IRT.

Analyses. In order to account for the potential difference in difficulty between the different exams each year, the scores on the three exams were equated using the 2-parameter logistic model (2PL) with a non-equivalent internal anchor test design (NEAT; Kolen & Brennan, 2014). Based on several simulation studies, Kolen and Brennan (2014) recommended using separate calibration rather than concurrent calibration for the NEAT design, in particular when the data do not fit the IRT model perfectly. Furthermore, in order to apply the separate calibration it is typically advised to have at least a 20% overlap between tests to be equated when the total number of items is 40 or more (Kolen & Brennan, 2014, p. 288). Table 4.4 shows that this percentage is met when equating the tests of cohorts 2013 and 2014 onto the test of cohort 2015.

Typically, the newer forms of tests are equated to the oldest form of a test, which would imply equating the tests of 2015 and 2014 onto the reference test of 2013. There were, however, only 11 anchor items between the exams of 2013 and 2014, which is less than the advised 20%. Two approaches could be taken to resolve this issue: equate 2015

onto 2013, and indirectly equate 2014 onto 2013 by the path 2014-2015-2013. Another approach, however, is to equate the tests of 2013 and 2014 onto 2015. We examined the results of both approaches, and found no substantial differences. Therefore, we report the results of equating test forms 2013 and 2014 onto the reference test of 2015, since the exam of 2015 was specifically designed to include items from both previous cohorts.

Based on the above considerations, the analyses proceeded as follows in *R*: first the 2-parameter logistic model (2PL) was fit to each data set separately and item-fit was evaluated using the ltm package (Rizopoulos, 2006, version 1.0). Subsequently, the Stocking-Lord test characteristic curve transformation method was used to transform the IRT scales of the 2013 and 2014 exams to the scale of 2015 using the package equateIRT (Battauz, 2015, version 2.0-3). In particular, this provided us with all estimated abilities on the same scale, as desired. The expectation of implementing practice tests in 2014 and 2015, was to improve student performance. Therefore, we used a one-sided test for the null-hypothesis that the 2014 and 2015 cohorts did not perform better than the 2013 cohort.

4.3.2 Results

Table 4.4 shows the equating coefficients and their standard errors for each separate calibration. The density curves of the equated ability, shown in Figure 4.6, are rather similar to each other with a vertical line showing the mean ability of students after the scores were equated. We tested the one-sided null hypothesis that the difference in mean ability of the cohort of 2014 and 2015 is not greater than the mean ability of the 2013 cohort (no practice tests implemented) to answer the research question. Table 4.5 shows that this one-sided null hypothesis could not be rejected.

Table 4.4. Number of students, number of test items, number of common items with the reference cohort 2015, the estimated equating coefficients (slope *A*, intercept *B*), and mean estimated ability after equating for each cohort

	***N* students**	***N* items**	***N* common items (%)**	***A* (*SE*)**	***B* (*SE*)**	**Mean ability (*SD*)**
2015	384	80				-0.02 (0.92)
2014	348	80	41 (51%)	0.83 (0.05)	-0.07 (0.07)	-0.11 (0.74)
2013	349	79	36 (46%)	0.88 (0.06)	0.09 (0.08)	0.05 (0.81)

Table 4.5. Independent samples *t*-tests, corresponding *p*-values, effect sizes and confidence interval around the effect sizes for differences in mean cohort ability

	***t*(*df*)**	***p*-value[a]**	**Cohen's *d***	**95% CI**
INT-2015 with INT-2013	-1.12 (730.34)	.87	-0.08	[-0.23; 0.06]
INT-2014 with INT-2013	-2.70 (690.70)	> .99	-0.20	[-0.35; -0.06]

[a]*p*-value represent the one-sided hypothesis that the cohort with practice tests performs better than the control cohort 2013.

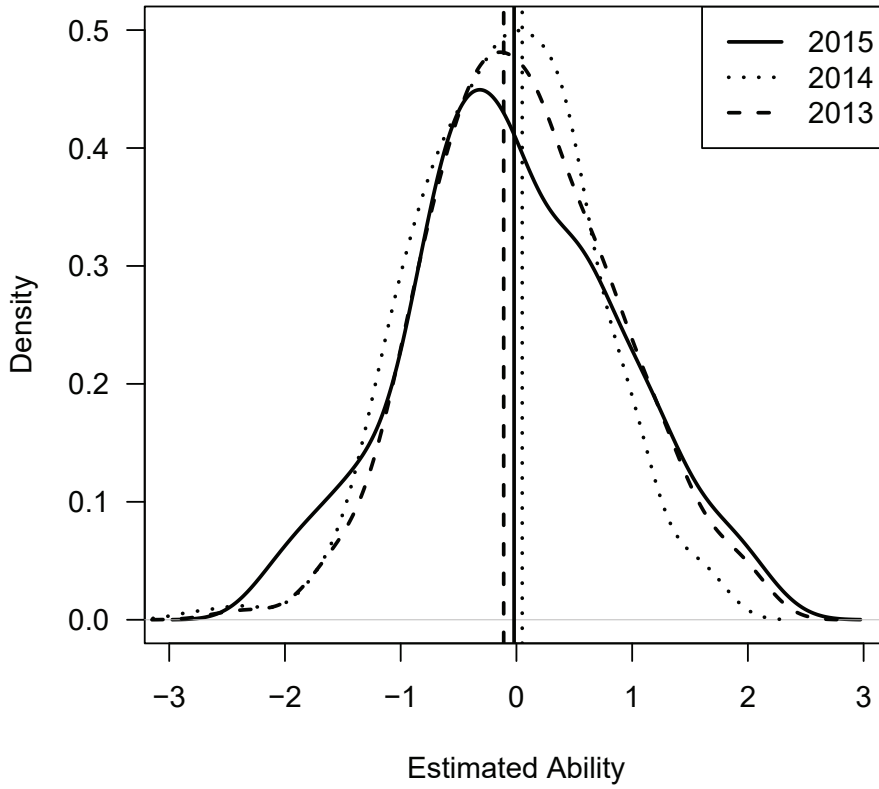


Figure 4.6. Density plot of estimated student ability for each cohort, with a vertical line at the mean ability.

4.4 Discussion

The aim of this study was twofold. First in Study 1 we sought to gain more insight into students' use of practice test resources, and the extent to which student's use of different types of practice tests was related to students' performance on the final exam in different courses with slightly different implementations of practice tests. The number of completed quizzes in Statistics 1b explained about 8% of the variance in student performance. Furthermore, we found about 5% explained variance in student performance in Statistics 1a, and a substantial amount of explained variance in Biopsychology (23%). While in all courses there was a positive relationship between the use of practice test resources and student performance, the strength of the relationship was greatest in the Biopsychology course with completely voluntary practice tests. In contrast to previous research, the incentives for using the practice tests in the statistics courses did not entail points that counted towards the final grade, but were part of a prerequisite in order to be able to complete the exam. These results seem to be in line with research demonstrating that providing incentives for use may encourage a use of tests that does not support the learning process, as suggested by Kibble (2007). Alternatively, not providing any incentives may reduce participation.

The lecturer of the statistics courses had mentioned the practice tests in the portal that accompanied the textbook, but found that these resources were not voluntarily accessed at all in previous cohorts. Thus, selecting the practice tests for students, and adding an incentive successfully encouraged their use to some extent, but not to a sufficient degree that its use was related to performance on the final exam.

In the second study we examined the extent to which cohorts in which practice tests were implemented performed better than a cohort where practice tests were not implemented. After equating the test scores, we found that the cohorts with access to practice tests did not perform better compared to the cohort with no practice tests. Thus, although there may be a positive relationship between the use of practice test and student performance (as illustrated in study 1), this does not directly translate to better student performance at the cohort level (as illustrated in Study 2).

A limitation of the present study, as in any research in educational practice, is that a teacher's goal to improve the learning process by offering practice tests cannot be causally verified to lead to better student performance. While this may theoretically be the case and a causal relationship may have been demonstrated in experimental research, it remains unclear which mechanism(s) may underlie the differential relationship between implementing practice tests and student performance. For example, student motivation and study skills have been found to explain 10-15% in student GPA (Robbins et al. 2004, Richardson et al., 2012), which is less than what was found in the present study for the Biopsychology course, but more than was found for the two statistics courses. Further research could investigate how practice test use could influence motivation and how this in turn affects subsequent use of practice tests, suggested by research on student perceptions of frequent assessment (Vaessen et al., 2016).

4.5 Conclusion

Offering practice test resources in higher education using web-based technologies is a way in which teachers can provide the means for students to receive immediate and individual feedback despite a small teacher-student ratio. Prior research, however, has used grading incentives and thus increased summative functions of practice resources (Kibble, 2007; Angus & Watson, 2009; Carrillo-de-la-Peña et al. 2009). In the present study, students' performance on the practice quizzes did not count towards the final grade, and students were given different types of resources to choose from at their own discretion.

Finally, we should realize that the relationship between the use of practice tests and study results may be a complex one. Good students may not use practice tests because they need less support in their learning process as was also found by research with primary school children (Roediger et al., 2011). On the other hand, students who do not use practice tests may lack the study skills to regulate their learning process and thus are unable to use the feedback from practice tests in a way that leads to better performance. Therefore, more insight is needed into how students' study behaviour is related to the relationship between practice test use and student performance.