U.PORTO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR
UNIVERSIDADE DO PORTO

The Effect of Evaluation on Learning

André Filipe Maia Silva

DISSERTAÇÃO

Unidade curricular "Dissertação/Projecto/Relatório de Estágio" do

6º ano do Mestrado Integrado em Medicina

Trabalho orientado por: Professor Doutor Manuel João Costa

Co-orientado por: Professor Doutor António Sousa Pereira

Junho 2011

ÍNDICE

ABSTRACT

In many undergraduate medical courses, tests are essentially used with the purpose of summative assessment, thus permitting teachers to measure student learning and to assign grades. However, research has shown that tests can also directly affect learning by promoting better retention of information, a phenomenon known as “testing-enhanced learning”.

The present research investigated whether test-enhanced learning can be effective in anatomical knowledge retention. Participants – first year medical students taking their first anatomy course - were divided in a test and a control group. The test group was submitted to two tests (in a period of 5 weeks) and the control group was only submitted to the last test. Both tests were designed according to the same blueprint and contained the same number of items. The content covered in the test had been presented in classes prior to the application of the first test. The difference between the two tests was that they were made of different items. We compared the results of test 1 and test 2 in the control group and also the results of test 2 between the test group and the control group.

Students in the test group had a significantly better performance in test 2 than in test 1; however, there was no significant difference between the results of the test group and the control group in test 2. This indicates that students improved their anatomical knowledge between the first and second test, but the increase cannot be directly attributed to the testing effect.

RESUMO

Em muitos cursos médicos pré-graduados, os testes de avaliação são utilizados para avaliação sumativa, permitindo aos professores avaliar a aprendizagem dos estudantes e dar notas. No entanto, a investigação científica mostrou que os testes de avaliação podem afectar directamente a aprendizagem, promovendo melhor retenção de informação, um fenómeno conhecido como aprendizagem potenciada pela avaliação.

O presente estudo investiga se a aprendizagem potenciada pela avaliação pode ser eficaz na retenção de conhecimentos de anatomia. Os participantes – estudantes do primeiro ano de medicina frequentando a primeira disciplina de anatomia do curso – foram divididos num grupo teste e num grupo controlo. O grupo teste foi submetido a dois testes (num período de 5 semanas) e o grupo controlo foi submetido apenas ao último teste. Ambos os testes foram construidos de acordo com a mesma matriz e continham o mesmo número de ítens. O contexto, objectivos e metodologia do projecto foram apresentados na aula de anatomia antes da aplicação do primeiro teste. A diferença entre os dois testes foi o facto de terem ítens diferentes. Comparámos os resultados do teste 1 e do teste 2 no grupo controlo e também os resultados do teste 2 entre o grupo teste e o grupo controlo.

Os estudantes do grupo teste tiveram resultados melhores no teste 2 do que no teste 1; no entanto, a diferença não foi significativa entre os resultados do grupo teste e do grupo controlo no teste 2. Isto indica que os estudantes melhoraram o seu conhecimento em anatomia entre o primeiro e o segundo teste, mas essa melhoria não pode ser directamente atribuida ao efeito do teste.

INTRODUCTION

Medical students study using different approaches. Some prefer repeated reading or re-writing, while others rather build concept maps, looking for connections between the studied materials. The efficacy of any approach depends on the students' characteristics and on the studied content. Teachers usually do not interfere with this process, letting each student use the approach of their preference. There are ways in which teachers can catalyze the learning process, like promoting frequent assessment[1,2].

The assessment of student learning is one of the most powerful educational tools[3]. Assessments can be formative or summative. Summative assessments are meant to measure what students learn for the purpose of making decisions - for example, pass/fail. Formative assessments are meant to catalyze the process of learning. Formative assessments are very important in the sense that they provide the individual student with feed-back concerning their strengths and weaknesses in a low-stakes context[4]. Students are thus more comfortable in taking risks and exposing what they do not grasp entirely.

Formative tests have the unexpected effect of enhancing memorization of factual content and skills. The importance of test-enhanced learning has been demonstrated on foreign language vocabulary learning[5], science education[1], art history[6] and resuscitating skills learning[7]. Larsen[8] concludes that tests that call for effortful recall of medical content are more effective in producing the retention of knowledge than repeated reading with no testing. Interestingly, the effect is stronger when feedback is provided after the test. Karpicke[5] compares the effect of testing to repetitive studying. The study reaches the conclusion that testing is more effective in promoting the memorization of factual content. Such studies have been replicated[1,2] thus indicating that, in medical education, testing can fulfill an important mission that is to help students memorize what they have studied, having an important influence in the learning process.

Anatomy teaching is constantly under modernization. New methods are continuously developed and test enhanced learning could be one of those[9].

The goal of the present study is to determine the effect of evaluation/ assessment on the anatomical knowledge. The underlying hypothesis was that students who would do a test on a certain anatomy content would achieve higher scores in a second test on the same content than students who had not taken the first test.

MATERIAL AND METHODS

The study was conducted in 7 weeks, approximately in the middle of the second semester in which the subjects are taking "Anatomy I", the first course of anatomy in the study plan (5,5 ECTS), which focuses essentially on human descriptive anatomy. The following information was solicited: birthdate and gender.

SAMPLE

The number of participants was 92 medical first year undergraduate students from one Portuguese medical school (35 in the test group and 57 in the control group). There were 4 students from the control group that dropped out of the study. The resulting study sample consisted of 88, 25 males (28%) and 63 females (72%). The median age of participants was 19,11 ± 2,59. The test group consisted of 31 students and the control group consisted of 57 students.

INSTRUMENTS

The work used two different short answer tests (APPENDIX II and III) developed from the same blueprint, with the same number of questions and content. (APPENDIX I e II) The questions were designed by an Anatomy Professor from another medical school and were considered adequate for application by a teacher of Anatomy from the institution where the study took place.

The blueprint of both tests consisted of:

Content	Items
Spacial relation of the head bones	4
The relation of the nerves, arteries and veins with the bones' foramina	6

The quality of the test was assessed calculating the internal consistency, difficulty indices and discrimination indices.

Each test contained 6 questions, written in the short answer format, which solicited 1, 2 or 4 elements for information totaling 10 possible correct/wrong answers. The students had 15 minutes to complete the test. The answers from students were compared with a pre-designed answer key with the 10 correct answers. The scorer in every answer weighed identically in the final score and the maximum possible total score was 10 points. In all circumstances that a student answer contained the key answer the total score was awarded. If the answer was incomplete, it was given the classification of 0,5 points.

PROCEDURES

The tests were administered in the context of a course of anatomy. The student learning outcomes defined for the course are to understand the structure and morphology of the human body in what

concerns osteology, myology and arthrology.

The project was presented to students twice, in the beginning of two theoretical anatomy classes in the same week, in order to maximize the number of students in the test group. The presentation explained the context, objectives and methodology for this study. Students were invited to enroll in the study by taking the first test immediately after the class. Such students were included in the test group and consist of 16 and 19 that took test 1, totaling 35 students.

The second test was administered four weeks later. There were 84 participants in the second test, among which 27 from the test group and 57 to the control group.

8 students dropped out of the study between the two administrations, 4 of which took the second test - two weeks after test 2 - after receiving a friendly email reminder. The results of the 4 students that only took test 1 were excluded from the analysis.

The final sample totaled 88 students, 31 in the test and 57 in the control group.

STATISTICS

For the statistical analysis we used Statistical Package for the Social Sciences (SPSS) 17.0 for Windows. For descriptive measures (absolute frequency, relative frequency, mean, standard deviation, maximum and minimum) we used a univariate analysis. A bivariate analysis was used to test differences between groups. To test normality, we applied the kolmogorov-smirnov test. T-tests were used to compare mean scores between groups.

ETHICS

This work was considered exempt from ethics review.

RESULTS

Internal consistency

The internal consistency of the test, Cronbach's alpha, was calculated to 0,634 in test 1 and 0,667 in test 2.

We analyzed the difficulty indices (APPENDIX I - tables 1 and 2) and discrimination indices (APPENDIX I - tables 3 and 4) of each question in both tests.

Testing the effect on learning

In order to identify the differences between the knowledge retention in the two groups, we calculated the descriptive statistics. (APPENDIX I - table 5 and 6).

For all statistical analysis we considered statistically significant p values < 0,05.

Normality

The distribution of the scores for the test group in tests 1 and 2 passed the normality test of Kolmogorov-Smirnov, but the control group in test 2 did not. However, parametric tests can be used in this sample, since the values for skewness and kurtosis for the

control group in test 2 were relatively low and their ratio to the standard error is <1.96.

To compare the differences between group scores in test 2, we used the t-test for independent samples. There was not a significant difference between the score from the test group to the control group, p-value: 0,63 (APPENDIX I - table 7).

To compare the differences between test 1 and test 2 in the test group, we used the t-test for paired samples. The score in test 2 was significantly higher than in test 1, p-value: <0,01 (APPENDIX I - table 8).

DISCUSSION

This study aimed at evaluating the effect of repeated testing on anatomical knowledge retention.

Test-enhanced learning has been demonstrated in various areas, medical[7,10] and non-medical[1,5,6] but not in the context of a topic traditionally difficult for students as anatomy. Anatomy requires extensive memorization, is very specific and less influenced by possible previous knowledge, eliminating the bias of knowing the content being evaluated before taking the course. The fact that these students are in their second semester and do not have previous related knowledge interfering with the learning process is also a positive factor towards choosing this curricular unit to the study. A positive and generalizable result would suggest that submitting students frequently to anatomy tests would be an effective way of promoting learning.

There was not a significant score difference between students who took a similar formative anatomy test twice and those who only took the second test. This was contrary to our original expectations. The results did however reveal that the subjects that took the two tests performed significantly better in the second. The fact that the two tests were constructed at the same time following the same blueprint guarantees that the variation covered is not biased due to inconsistencies between the tests. It is not possible to exclude the possibility that some students might have memorized items, but that is unlikely since the tests were low stakes and 30 days elapsed between the two administrations. Therefore, the differences in test scores should reflect actual variations in knowledge between the two moments.

The participants did not receive any feed-back on their answers between the two tests, as a means to restricting the reasons for variations in scores to the testing moment since feedback has a positive effect on retention[2,7].

Stress can influence the learning process[11,13]. Summative assessment has a stress generator effect that could possibly enhance the anatomical knowledge retention, thus potentiating the testing effect. However, we decided to create a

formative assessment in order not to interfere with the anatomy curricular program and to avoid specific preparation from students, so that we could restrict the external influences and assess only the testing effect. It is documented that formative assessment as a positive effect on learning[4].

Special care was taken in the construction of the test, taking into account recent references in what concerns assessment[12]. We chose to use a short answer test because it is considered superior to multiple choice questions in promoting long term knowledge retention[6]. The discrimination indices of 70% of the questions in each test were high, confirming that the tests were able to adequately discriminate students according to anatomical knowledge. Although the tests had a small number of questions, due to time constraints regarding the students, the Cronbach alpha, was above 0,6 in both tests. The internal consistency of tests improves with the size of the tests.

It is important to analyze the methodology, to identify confounding factors that may have influenced the results. A generalizable result would call for the use of a rigorous research methodology. The study was originally designed to include two groups, with the same number of students, randomized according to gender, administered information included in medical school application and socio demographic information. One of those groups would take 2 tests and the other would only take the second test. The tests would be applied in the end of the anatomy theoretical classes.

As is often the case when research is developed in real world scenarios, original experimental designs suffer adaptations due to constraints of the subjects and the context of the research itself.

Despite the efforts to promote the study in the student population, an insufficient number of participants dictated important deviations to the original study design. Therefore, two opportunities to take the first test were offered in the same week, immediately after theoretical anatomy classes. We cannot exclude possible communication about the test content between students who did the first test in the first opportunity and those who did it in the second opportunity. The second test was administered 4 weeks later, but this time, during another anatomy class, which increased the number of participants in the study. The number of students who volunteered to participate in the study was insufficient to conduct the randomization originally planned; therefore, all the students available to take the first test were allocated to do both tests. Because the test group sample was small and there were few men, we did not compare score differences between genders, although some differences were recorded in a previous

study[11]. One further important limitation was that there was a summative assessment 2 weeks after the first test and 2 weeks before the second test which may have influenced the results and the testing effect.

Further investigation is needed to clarify the positive effect of formative test enhanced learning in anatomy teaching. This study did not demonstrate such effect, but obtained positive results in circumstances of important constraints to the experimental design. It would be interesting to replicate the study with the correct design, to investigate the "test-enhanced learning" in anatomy education.

ACKNOWLEDGEMENTS

I wish to thank: Manuel João Costa, PhD, for all the constant support and the enormous amount of time spent in the construction of this research and writing review; António Sousa Pereira, PhD, for the good discussions and important inputs; João Cerqueira, PhD, for the test construction and anatomical knowledge contributions; Artur Águas, PhD, for allowing the research during the anatomy classes period and helping in the test application logistics; Ana Rita Lopes, for the last minute study review; all the study participants that spent their precious time taking the tests.

BIBLIOGRAPHY

1 Karpicke JD, Blunt JR. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. Science 2011; 331(6018):772-5

2 Butler A C. Repeated Testing Produces Superior Transfer of Learning Relative to Repeated Studying Journal of Experimental Psychology: Learning, Memory and Cognition 2010; 36 (5): 1118–1133

3 Wormald BW, Schoeman S, Somasunderam A, Penn M. Assessment drives learning: an unavoidable truth? Anat Sci Educ 2009; Oct;2(5):199-204.

4 Carrillo-de-la-Peña MT, Baillès E, Caseras X, Martínez A, Ortet G, Pérez J. Formative assessment and academic achievement in pre-graduate students of health sciences. Adv Health Sci Educ Theory Pract 2009; 14(1):61-7.

5 J. D. Karpicke, H. L. Roediger. The Critical Importance of Retrieval for Learning. Science 2008, 15, 966-968.

6 Butler AC, Roediger, HL, III. Testing improves long-term retention in a simulated classroom setting. European Journal of Cognitive Psychology 2007, 19 (4/5), 514-527

7 Kromann C, Jensen ML, Ringsted C. The testing effect in skills learning. Med Educ 2009; 43:21–7.

8 Larsen, D P, Butler, A C, Roediger, H. L, III. Repeated testing improves long-term retention relative to repeated study: A randomized, controlled trial. Medical Education 2009, 43, 1174-1181

9 Sugand K, Abrahams P, Khurana A. The anatomy of anatomy: a review for its modernization. Anat Sci Educ. 2010; 3(2):83-93.

10 Larsen DP, Butler AC, Roediger HL, III. Test-enhanced learning in medical education. Med Educ 2008; 42:959–66

11 Kromann CB, Jensen ML, Ringsted C. Test-enhanced learning may be a gender-related phenomenon explained by changes in cortisol level. Med Educ 2011; 45(2): 192-9.

12 Norcini J, Anderson B, Bollela V, Burch V, Joao Costa M, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, Roberts T. Criteria for Good Assessment: Consensus Statement and Recommendations; Medical Teacher 2011; 33: 206-214.

13 De Maria S, Bryson EO, Mooney TJ, Silverstein J, Reich DL, Bodian C, Levine AI. Adding emotional stressors to training in simulated cardiopulmonary arrest enhances participant performance. Med Educ 2010; 44: 1006–1015.

APPENDIX I

Table 1 – Difficulty indices – Test 1

A1	A2	B1	B2	B3	B4	C	D	E1	E2
0,935	0,032	0,323	0,355	0,290	0,193	0,226	0,452	0,581	0,129

Table 2 – Difficulty indices – Test 2

A1	A2	B1	B2	B3	B4	C	D	E1	E2
0,977	0,00	0,648	0,716	0,682	0,364	0,534	0,023	0,841	0,625

Table 3 – Discrimination Indices – Test 1

A1	A2	B1	B2	B3	B4	C	D	E1	E2
0,250	0,00	0,500	0,750	1,00	0,625	0,562	0,125	0,812	0,312

Table 4 – Discrimination Indices – Test 2

A1	A2	B1	B2	B3	B4	C	D	E1	E2
0,00	0,00	1,00	0,733	1,00	0,633	0,633	0,067	0,533	0,500

Table 5 - Descriptive Statistics – Test 2

	grupo			Statistic	Std. Error
t2total	teste	Mean		5.6774	.34745
		95% Confidence Interval for Mean	Lower Bound	4.9678	
			Upper Bound	6.3870	
		5% Trimmed Mean		5.7527	
		Median		6.0000	
		Variance		3.742	
		Std. Deviation		1.93455	
		Minimum		2.00	
		Maximum		8.00	
		Range		6.00	
		Interquartile Range		3.00	
		Skewness		-.576	.421
		Kurtosis		-.865	.821
	controlo	Mean		5.8728	.23019
		95% Confidence Interval for Mean	Lower Bound	5.4117	
			Upper Bound	6.3339	
		5% Trimmed Mean		5.9476	
		Median		6.5000	
		Variance		3.020	
		Std. Deviation		1.73792	
		Minimum		2.00	
		Maximum		8.00	
		Range		6.00	
		Interquartile Range		3.00	
		Skewness		-.534	.316
		Kurtosis		-.932	.623

Table 6 – Descriptive Statistics – Test 1

			Statistic	Std. Error
t1total	Mean		3,7258	,35815
	95% Confidence Interval for Mean	Lower Bound	2,9944	
		Upper Bound	4,4573	
	5% Trimmed Mean		3,6855	
	Median		3,5000	
	Variance		3,976	
	Std. Deviation		1,99411	
	Minimum		,00	
	Maximum		8,00	
	Range		8,00	
	Interquartile Range		3,00	
	Skewness		,248	,421
	Kurtosis		-,495	,821

Table 7 - Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
t2total	Equal variances assumed	,441	,508	-,484	86	,630	-,19539	,40369	-,99790	,60712
	Equal variances not assumed			-,469	56,304	,641	-,19539	,41679	-1,03022	,63944

Table 8 - Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	t2total - t1total	1,95161	2,27525	,40865	1,11705	2,78618	4,776	30	,000

APPENDIX II

Código: _____

A duração máxima do teste é de 15 minutos. Cada pergunta deve ter uma resposta curta.

A. Mulher de 52 anos trazida ao Serviço de Urgência na sequência de um acidente de automóvel. Apresenta hematoma na testa. Assumindo este como o único local de impacto, em que ossos é mais provável encontrar fracturas?

Frontal e occipital_____

B. Homem de 28 anos, trabalhador da construção civil, trazido ao Serviço de Urgência após queda em altura de objecto pesado sobre o crânio, não protegido com capacete. Sem lacerações da pele. Verifica-se um hematoma na fissura orbital superior. Que nervos poderão ter sido afectados?

Oftálmico, oculomotor, troclear, abducente_____

C. Homem de 33 anos, a quem é diagnosticada uma fractura da parede lateral dos seios etmoidais direitos. Com que cavidade passaram os seios a comunicar?

Órbita_____

D. Que osso separa a sela turca da cavidade nasal?

Esfenóide_____

E. Homem de 16 anos, com fraqueza em todos os movimentos da mandíbula dois dias após ter sofrido traumatismo craniano enquanto jogava basquetebol. O médico explicou-lhe que tal se devia a uma lesão do nervo que atravessa o buraco oval. De que nervo se trata?

Nervo mandibular_____

Se o buraco atingido fosse o espinhoso, que estrutura vascular poderia ser atingida?

Artéria Meníngea Média_____

APPENDIX III

Código: _____

A duração máxima do teste é de 15 minutos. Cada pergunta deve ter uma resposta curta.

A. Homem de 50 anos, trazido ao Serviço de Urgência na sequência de acidente de motorizada, com queda para trás. Apresenta hematoma occipital. Assumindo este como o único local de impacto, em que ossos é mais provável encontrar fracturas?

Occipital e frontal_____

B. Mulher de 23 anos, trazida ao Serviço de Urgência após queda do 3º andar. Sem lacerações da pele. Verifica-se um hematoma no buraco jugular. Que estruturas nervosas e vasculares poderão ter sido afectadas?

Nervos glossofaríngeo, vago, acessório e veia jugular interna_____ _____

C. Mulher de 32 anos, a quem é diagnosticada uma fractura do pavimento da órbita. Onde é de esperar encontrar sangue depositado?

Seio maxilar_____

D. Que osso separa o andar anterior do crânio da cavidade nasal?

Frontal_____

E. Mulher de 15 anos, com fraqueza nos movimentos da face dois dias após ter sofrido traumatismo craniano enquanto escalava. O médico explicou-lhe que tal se devia a uma lesão do nervo que atravessa o buraco estilomastoideu.

1. De que nervo se trata?

Facial_____

2. Se o buraco atingido fosse o carotídeo, que estrutura vascular poderia ser atingida?

Artéria carótida interna_____