Universidade do Minho
Escola de Psicologia

Pedro Simão Oliveira Mendes

Testing effect on a college course: Examining test format and repeated questioning

junho de 2015

Universidade do Minho
Escola de Psicologia

Pedro Simão Oliveira Mendes

Testing effect on a college course: Examining test format and repeated questioning

Dissertação de Mestrado
Mestrado Integrado em Psicologia

Trabalho efetuado sob a orientação do
Professor Doutor Pedro B. Albuquerque
e coorientação da
Professora Doutora Véronique Quaglino

junho de 2015

DECLARAÇÃO

Nome: Pedro Simão Oliveira Mendes

Endereço eletrónico: pso.mendes@gmail.com

Número do Cartão de Cidadão: 13951590

Título da dissertação: **Testing effect on a college course: Examining test format and repeated questioning**

Orientador: Professor Doutor Pedro B. Albuquerque

Co-Orientador: Professora Doutora Véronique Quaglino

Ano de conclusão: 2015

Designação do Mestrado: Mestrado Integrado em Psicologia

Universidade do Minho, 12/06/2015

Assinatura: _____

Contents

Agradecimentos

Ao professor doutor Pedro B. Albuquerque, pela oportunidade de desenvolver um trabalho numa área do meu interesse, pelo apoio e pela orientação constantes, pela exigência na realização de um bom trabalho. Também pelo seu talento no ensino, que desde logo me fez interessar pela temática da Memória Humana.

À professora doutora Véronique Quaglino, pelo seu feedback interessantíssimo.

À minha família. Aos meus pais, pelo seu esforço incansável no investimento da minha formação. À minha irmã por todo o apoio ao longo destes cinco anos. À minha avó, tia-avó e tias. À tia Irene, à tia Ana Maria, à D. Fernanda. Sem vocês este percurso não teria sido possível.

Ao Sr. Francisco e à D. Virgínia, que foram o meu apoio nos momentos menos bons e me ajudaram a levantar.

À Cátia Marisa Graça, um obrigado especial pela sua presença em todos os momentos desta longa caminhada. Pelo seu amor e carinho. Contigo a meu lado, tudo foi, tudo é e tudo será mais fácil.

A todos os meus colegas e amigos que, de uma maneira ou de outra, ao longo destes cinco anos, influenciaram o meu percurso e a minha vida. Um obrigado especial à Inês, à Margarida, à Bruna, à Cátia, à Joana, ao Francisco, e ao Israel, os *Psis da Vida Airada*. Um obrigado especial também a quem me acompanhou na aventura em Lille: Daniela, Johanna e Rui. Um obrigado especial também ao André Silva, pela ajuda essencial na análise de dados. E por fim, a todos os que me acompanharam nas diferentes equipas da AEPUM, pela amizade e pelas experiências de aprendizagem que me proporcionaram.

Efeito de testagem em contexto universitário: Estudo da influência do formato de teste e questionamento repetido

Os testes parecem promover a retenção a longo-prazo, fenómeno conhecido como o efeito de testagem. Face à robustez do efeito tem crescido o interesse em transferir o efeito obtido em laboratório para o contexto educacional. Sabe-se que o formato do teste (escolha múltipla ou completamento de frases) modula o tamanho do efeito e que a testagem pode promover a retenção de matéria relacionada. Este estudo pretendeu examinar o papel da testagem repetida através de minitestes semanais no desempenho académico dos alunos numa frequência. Setenta e cinco estudantes universitários responderam a 6 minitestes semanais, cada um abrangendo os tópicos dum capítulo do manual. Então, realizaram uma frequência que avaliava todos os tópicos, através de questões Repetidas, Relacionadas e Novas. Os resultados mostraram que os alunos responderam melhor às perguntas Repetidas (M =.72) e Relacionadas (M =.71) do que às Novas (M =.61). Questões de escolha múltipla nos minitestes produziram maior efeito de testagem na frequência do que as de completamento. Os resultados suportam as evidências de que o efeito de testagem se aplica ao contexto educacional, não só para questões repetidas, mas também para questões relacionadas, em comparação com questões novas.

Palavras-chave: Efeito de testagem, educação, testes, facilitação

Testing effect on a college course: Examining test format and repeated questioning

Testing seems to benefit long-term retention, a phenomenon known as the testing effect. There is a growing interest on transferring the testing effect to the classroom. Test format [multiple-choice (MC) or short-answer (SA)] is known to modify the enhancement of testing, and testing may also promote retention of non-tested material. This study aimed to examine the role of repeated testing through weekly quizzes in student's performance on an intermediate test with tested and related non-tested questions, when comparing them to new questions. We also wanted to clarify the effect of question format on later retention. 75 college students answered 6 weekly quizzes comprising the topics of a book chapter each. Then, they had an intermediate test about all the 6 topics, with Repeated (RP), Related (RL), and New (N) questions). Overall results showed that students' performance was better for RP (M= .72) and for RL (M= .71) than for N (M= .61). When quizzes were in MC format, testing effect was larger than if SA format. Results support not only the evidence that the testing effect is transferable to the classroom, but also that testing may benefit student's performance both on tested and related questions when compared to new questions.

Keywords: Testing effect, retrieval-induced facilitation, education, quiz

Testing effect on a college course: Examining test format and repeated questioning

The *testing effect* refers to the enhancement of later retention after engaging on retrieval practice, or testing (Roediger & Karpicke, 2006b). Many studies have found evidence for the testing effect with different materials: pictures (Wheeler & Roediger, 1992), paired-associated words (Jacoby, 1978) and more recently with educational-relevant material, like prose passages (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Roediger & Karpicke, 2006a; Roediger & Marsh, 2005) or definitions of concepts (Rawson, Dunlosky, & Sciartelli, 2013) and even cardio-pulmonary resuscitation skills (Kromann, Bohnstedt, Jensen, & Ringsted, 2010). In fact, it seems that the testing effect occurs independently of the type of material learned (Rowland, 2014). Through the past decade, there has been a growing interest on the testing effect, both in the laboratory and in applied contexts (e.g., classrooms), with different approaches. In fact, the phenomenon has been given so much attention that there was a recently interest in understanding its processes through studies that explore its neural correlates (van den Broek, Takashima, Segers, Fernández, & Verhoeven, 2013; Wing, Marsh, & Cabeza, 2013).

Testing effect in educational settings

Effective long-term learning is the main goal of education and a lifelong pursuit. Studies with educational-relevant material have obvious practical implications regarding the enhanced learning from testing (e.g., how many test students should take). Despite the growing interest on this topic, there are still few studies examining the testing effect in educational settings (McDaniel, Roediger, & McDermott, 2007). Nevertheless, the finding that the testing effect is applicable to classroom context seems too convincing to be unconsidered. Although most participants of these studies are young adults or college students (McDaniel et al., 2007; Glass, Ingate, & Sinha, 2013), there are also evidence for testing effect benefits with middle-school students (McDaniel et al., 2013) and older adults (Meyer & Logan, 2013). An interesting finding is that students seem to be unaware of the testing effect and predict they will be more successful when engaging on repeated study (Agarwal et al., 2008; Roediger & Karpicke, 2006a). In fact, most students seem to engage in repeated reading rather than self-testing or retrieval practice while studying (Karpicke, Butler, &

Roediger, 2009). More, it seems that these habits tend to be carried out through life, suggesting that even away from the time-pressure caused by exams, people do not use the best strategies to enhance self-regulated learning (Yan, Thai, & Bjork, 2014).

Students are continually exposed to evaluations and they typically are graded through tests from kindergarten to college. Tests may enhance long-term retention directly or indirectly, either by enhancing students awareness of the contents they might know better or worse, allowing a more effective study, either by reducing test anxiety (Agarwal, D'Antonio, Roediger, McDermott, & McDaniel, 2014). Even so, tests usually only have a grading purpose (McDaniel et al., 2007; Roediger & Karpicke, 2006b; Roediger & Butler, 2011).

Findings on the testing effect in the classroom are known since the classic studies of Gates (1917) or Spitzer (1939), but it was assumed for a very long time that learning only occurred during study sessions. The results from Tulving (1967) study with lists of words showed that learning also occurred during tests, and recent findings replicated them with a more controlled procedure (Roediger & Butler, 2011; Roediger & Karpicke, 2006b). In general, when compared to a repeated-study condition (i.e. the participants read several times the material to be tested later), a repeated-test condition (i.e. the participants read the material once which is tested several times before a final test) enhances later retention of the material. Although they perform poorly on an immediate test (e.g., 5 min after study), participants perform better on a later test (e.g., 2 days or 1 week later) (Roediger & Karpicke, 2006b). The testing effect occurs both with between-subjects designs (like the previous example) and within-subjects designs. When within-subjects designs are used, participants are exposed to the material, but only part is tested, while other part is not tested (e.g., Kang, McDermott, & Roediger, 2007). Participants perform better in a long-term retention test for the materials that were previously tested. One might think that the testing effect occurs because participants are more exposed to the material during testing trials, justifying their better performance compared to a trial of repeated study. Rowland's recent meta-analysis (2014), based on studies with a restudy control condition (thus, time of exposure controlled), indicates the reliability of the testing effect and rules out the mere exposure effect as its source.

According to Roediger and Butler (2011), the increased retention caused by repeated retrieval seems to depend on several factors, such as the interval between successive retrievals or the number of successful retrievals. Spaced testing seems to have larger and longer-lasting effects on retention than massed testing within a session (Rawson et al., 2013). Even if there

is evidence suggesting retention enhancement by a single test (Carpenter & DeLosh, 2006), there is also research showing the benefits of repeated testing over single test (Roediger & Karpicke, 2006a). Curiously, the results from Rowland's meta-analysis (2014) showed that the size effect of a single test is similar with the one caused by multiple test. Also, it supports the idea that the testing effect is greater for longer retention intervals (larger magnitude effects), but suggests "testing reliably benefits retention compared with restudy even at short intervals" (p. 1449).

Question format

As Kang et al. (2007) point out, the type of questions is an important aspect to consider regarding the testing effect. For example, Roediger and Marsh (2005) used a total of 36 nonfiction passages that spanned a variety of topics (e.g., science, famous people, history, places and animals) to create multiple-choice (MC) questions with a variable number of options (0, 2, 4, 6), where only one answer was correct. College undergraduates read half of the passages (counterbalanced in two groups of 18 passages) and were tested in all of them (questions about non-read passages were control), then took a filler task before a final cued-recall test. Results showed that more questions were answered correctly when participants had read the relevant passages and when tested with fewer alternatives (on the MC tests). On the final cued recall test, there was a strong testing effect for read passages and there were more wrong answers on questions with more options, which might suggest that there may be a negative interference of the wrong answers when a MC question has many options (negative suggestion). This study makes note of possible negative effects of multiple-choice questions. Even so, according to Bjork, Little and Storm (2014), multiple-choice testing may prove to be a desirable difficulty in the classroom, as it enhances not only the retention of tested information, but also of conceptually similar information.

Kang et al. (2007) also examined whether multiple-choice (MC) or short-answer (SA) tests lead to better later retention. Their results (Exp. 1), showed that on a final test, even though both MC and SA tests lead to a testing effect, having a MC test as retrieval practice led to better performance than if the test was in SA format, but the MC retrieval practice condition did not differ significantly from the read condition. In a second experiment, to examine the effect of feedback after a testing trial on a final test, the authors included feedback after each question on the intervening tests. The results showed that taking an SA retrieval practice test with corrective feedback produced the best final test performance for

both MC and SA formats (Kang et al., 2007 – Exp. 2). These findings support the idea that recall tests are more beneficial for subsequent retention than recognition tests or additional study. McDaniel, Anderson, Derbish and Morrisette (2007) found similar evidence in an experiment conducted with students enrolled in an online course, comprising educational materials (40-page reading assignment from a textbook). They found that even in a variable environment such as the educational context, SA weekly quizzes produced a greater testing effect than MC weekly quizzes (both with immediate feedback).

Immediate feedback after retrieval practice (test or quizzes) may also nullify the interference of wrong answers given, helping students perform better on a final test (Butler & Roediger, 2008). In educational settings, however, students cannot always be provided immediate corrective feedback. Kornell (2014) examined the effects of delayed feedback on learning word pairs and more meaningful stimuli (e.g., trivia questions). The results suggest that the beneficial effects of feedback are found even when it is delayed up to one day.

Also, one might assume that the testing effect is similar to the generation effect (Slamecka & Graf, 1978) – generating items during an initial learning phase often enhances performance on a memory test, – considering that SA questions can be a generation task. Even so, there are differences between both effects, as generation requires retrieval from semantic memory but testing requires retrieval from episodic memory (Rowland, 2014), through intentional retrieval. Karpicke and Zaromb's (2010) manipulated the instruction given to participants (whether explicitly to retrieve items from a previously learned list; or to generate them from semantic memory). They found that explicit instruction to engage in intentional retrieval led to a better performance than the generation tasks on later retention, indicating a difference in the magnitude of the effect. Rowland (2014) states that the design of the study may influence the magnitude of both generation effect and testing effect, with the first appearing larger in within-subjects designs (see Bertsch, Pesta, Wiscott, & McDaniel, 2007) and the latter being larger in between-subjects designs.

Theories on the testing effect

Theoretical explanations for the testing effect have focused on the transfer-appropriate processing hypothesis and on the elaborative retrieval hypothesis (Roediger & Butler, 2011; Rowland, 2014). Transfer-appropriate processing states that memory performance is enhanced when the cognitive processes involved during retrieval (testing)

match those used at encoding (study) moment. The elaborative retrieval hypothesis posits that the testing effect occurs because an effortful retrieval of information leads to elaboration of the memory trace and/or creates additional retrieval routes or pathways, making the information more likely to be retrieved later. Results from different studies (Carpenter & DeLosh 2006; McDaniel et al., 2007; Kang et al., 2007) present evidence against the transfer-appropriate processing hypothesis, but in favor of the elaborative retrieval hypothesis.

For example, Carpenter and DeLosh (2006) explored whether the magnitude of the testing effect would reflect the match between retrieval practice tests format and final tests format. They compared matching (final test in the same format as the retrieval practice) and mismatching (final test in different format as the retrieval practice) conditions, founding that retention was not enhanced for matching, relative to mismatching, which is contrary to the transfer-appropriate–processing hypothesis (experiment 1). More, these authors controlled individual item difficulty and manipulated the number of cues present, finding that retrieval practice seems most beneficial to final retention when it provides more potential for elaborative processing (experiment 3).

In the meta-analysis conducted by Rowland (2014), "initial–final test match did not yield a reliable increase in the magnitude of the testing effect" (p. 1447) comparing to mismatching, suggesting once more that transfer-appropriate processing hypothesis may not be the best to explain the testing effect. Besides, the results seem to support the prediction that more effortful tests (recall relatively to recognition) should produce larger testing effects.

Testing effect on related material

Besides the effects of testing on tested material, there are findings showing that retrieval practice may also promote retention of related non-tested material - retrieval-induced facilitation (Chan, McDermott, & Roediger, 2006). Results also show an impairment on retention of related non-tested material - retrieval-induced forgetting (Chan, 2009, 2010). Chan (2009) tried to understand when testing did led to retrieval-induced forgetting and when did it let to retrieval-induced facilitation of related non-tested materials. He manipulated the level of integration invoked at encoding (refers to the level the tested items can be associated to the related non-tested items) and the length of delay between retrieval practice and the final test, two conditions that may affect these phenomena (e.g., high levels of integration could eliminate the retrieval-induced forgetting). The results suggested that when the final test

occurs after 24 hours, participants in a high integration condition may benefit of a retrieval-induced facilitation; but when it occurs after 20 minutes, participants in a short delay low integration can experience retrieval-induced forgetting. This information may be useful for students to increase the positive effect of retrieval practice (e.g., performing retrieval practice immediately before an exam should be avoided, especially if the to-be-tested materials are similar to the ones practiced) (for more details, see Chan, 2009). Chan (2010) also found that the retrieval-induced facilitation may be present after a long-term interval (7 days). However, Chan's studies were not conducted with authentic educational materials, as was the recent study by Wooldridge, Bugg, McDaniel, and Liu (2014). These authors found that although there were benefits for repeated information previously tested, topically related information did not benefit from the testing effect with educational materials. Even so, not even this study was conducted in genuine educational settings.

Extending the testing effect to educational settings can help improving students' long-term learning, since it can be a tool used by educators or students themselves (e.g., to guide their study). Hence, this study aims to examine the role of repeated testing in students' performance on an intermediate test with repeated tested questions as well as related non-tested questions. We intend also to clarify the effect of question type (MC vs. SA) on later retention in everyday educational settings. Students' performance should be better for tested questions than for non-tested questions, and the benefits of testing may also occur for related questions when comparing with non-tested questions. If no feedback is provided, MC questions are expected to produce a stronger testing effect than SA questions.

Method

Participants

Seventy-five students (60 female and 15 male) participated in this experiment. Students were enrolled on the first year course of Psychology of Memory, a course of the Graduation in Psychology (University of Minho, Portugal). Participants completed all of the evaluation components of the course and agreed to participate in the study.

Materials

Questions were constructed based on *Memory* (Baddeley, Eysenck, & Anderson, 2009 - Brazilian Portuguese translation), the adopted textbook for the course of Psychology of Memory. The course, Psychology of Memory, was organized in 14 weeks of contact with the students (classes), with an hour block on Mondays and two hour block on Thursdays. The course comprised 12 topics on Human Memory, based on 12 chapters of the adopted textbook.

The assessment methods for the course comprised 12 Quizzes (Q), 2 Intermediate Tests (IT) and 1 Final Exam (FE). Each Q comprised 10 questions about one topic, each IT comprised questions about six topics, and the FE aggregates questions about all of the 12 topics. Quizzes, Intermediate Tests and the Final Exam determined 30%, 30% and 40% of students' final grade on the course. This study took into account results from the first 6 Quizzes and the first Intermediate Test.

Each Quizz was 10 questions long, either multiple-choice (MC) or short-answer[1] (SA), and comprised a book chapter (e.g., Working Memory). The Intermediate Test was composed of 72 questions (both MC and SA), 12 per each chapter. From those 12 questions, 4 were Repeated questions (RP), 4 were Related questions (RL), and 4 were New questions (N). Questions consisted of a statement with a missing part, which students had to fill in, either writing it down on the blank space (SA questions – e.g., *In his study, Hermann Ebbinghaus (1885) used ___ as stimuli.*) or choosing the right answer to fill the blank space from four different given options (MC questions – e.g., *In his study, Hermann Ebbinghaus (1885) used ___ as stimuli. (a) words; (b) nonsense syllables; (c) monosyllables; (d) acronyms*).

The Repeated questions (2 MC and 2 SA per chapter) were chosen based on the following criterion: proportion of correct answers on Quizzes ranged between .49 and .80, so

[1] In this study, each question was a sentence with missing information (blank space). Students should complete the sentence with the missing concept. In this sense SA questions were sentence completion questions, where students must write the answer, filling the blank space (see example in text above)

that questions were not too easy (proportion of correct answers above .80), avoiding a ceiling effect, but also not too difficult (proportion of correct answers below .49), avoiding a floor effect. In order to guarantee that MC and SA Repeated questions had the same difficulty, the means of MC questions and SA questions chosen were compared. No differences, $t(11) = .42$, $p = .680$, between the means of MC ($M_{MC} = .66$, $SD = .10$) and SA questions ($M_{SA} = .65$, $SD = .09$) were found.

Table 1 - Example of type of question (correct answers in parenthesis)

	Type of question		
	Repeated	**Related**	**New**
Quiz	In his study, Herman Ebbinghaus (1885) used _____ as stimuli. (*nonsense syllables*)	According to the _____, the more time you spend studying a material, the more information you will store and retrieve. (*total time hypothesis*)	-
Intermediate Test	In his study, Herman Ebbinghaus (1885) used _____ as stimuli. (*nonsense syllables*)	If I want to get better on my writing skills by simply spending more time "writing", I'm founding this knowledge on the _____. (*total time hypothesis*)	The fact we know how to ride a bike, even when we don't do it often, can be explained because this knowledge results from a _____ learning. (*procedural*)

Design

A within-subjects design was used. Type of question (Repeated, Related, New), question format (Multiple-Choice, Short-Answer) and format congruency (Same format on Q and IT, Different format on Q and IT) were the independent variables manipulated.

The dependent variable was the proportion of correct answers on the questions presented in the Intermediate Test.

Procedure

After a theoretical class about each topic, students answered weekly quizzes, 10 questions long, comprising the same topic, with the interval between the class and the quiz being always three days. Question format was counterbalanced between two groups of students, A (N = 36) and B (N = 39), so that group A answered the Quizzes in this order: MC, SA, MC, SA, MC, and SA. Group B answered the same quizzes on the inverse order of group A. Students had 10 minutes to answer the Quiz.

The Intermediate Test was 72 questions long that could be Repeated, Related or New. From each of the 4 Repeated, Related or New questions there were 2 MC questions and 2 SA questions. In this way, for all students there were questions that were answered on the same format both on Quizz and Intermediate Test and questions that were answered on Different format, either MC on Quizz and SA on Intermediate Test, or SA on Quizz and MC on Intermediate Test. Students had 75 minutes to complete the Intermediate Test.

Although the questions format (MC or SA) might vary, Repeated questions were exactly the same, both on Quizzes and Intermediate Test. The Related questions were constructed based on the same unit of a specific topic that was assessed on the Quizzes, but the question *per se* was different (meaning they would evaluate the same concept, though the wording for the question was different). New questions were totally new questions, based on topics not assessed on the Quizzes, serving as baseline performance for students on each topic. On table 1 there is an example of questions for chapter 4, about Memory and Learning.

On each evaluation moment (Quiz and Intermediate Test) students were asked to predict how many questions they would answer correctly and to estimate the number of hours they had spent studying for that Quiz or Intermediate Test.

Results

Results focus on student's performance on the Intermediate Test, and consider the type of question (Related, Repeated and New) as well as the questions format. Question format congruency (Same or Different) was explored, as well as the accuracy for repeated questions (on Quizzes and Intermediate Test). All analysis were performed using Microsoft Excel 2013 and IBM Statistical Package for the Social Sciences (SPSS) software (version 22).

Students' performance on the Intermediate Test

On the Intermediate Test, students answered Repeated, Related and New questions, both on Multiple-choice format or Short-answer format. The mean proportions of correct answers for each type of question and question format are presented on Table 2.

Table 2 – Mean proportion of correct answers on Intermediate Test according to question format

	Repeated	**Related**	**New**
	M (SD)	*M (SD)*	*M (SD)*
MC	.76 (.16)	.77 (.13)	.80 (.13)
SA	.69 (.22)	.64 (.18)	.42 (.18)

Note: MC= Multiple-choice; SA= Short-Answer.

A 3 (Type of Question: Repeated, Related, New) x 2 (Question Format: Multiple-Choice, Short-Answer) within-subjects ANOVA was conducted. Overall, there was a main effect for type of question, $F(2, 73) = 49.24$, $p < .001$, $\eta_p^2 = .40$. The pairwise comparisons revealed that students performed better for Repeated questions ($M_{rep} = .72$, $SD = .18$) than for New questions ($M_{new} = .61$, $SD = .13$), $p < .001$, 95% CI [.08, .15]; and for Related questions ($M_{rel} = .71$, $SD = .14$) than for New questions, $p < .001$, 95% CI [.07, .13]. No differences were found between Repeated and Related questions, $p = .57$, 95% CI [-.01, .05].

Pairwise comparisons considering independently MC and SA questions, showed that within MC questions, no differences were found between students' performance for Repeated, Related or New questions ($p < .05$ in all conditions). However, within SA questions, students performed significantly better for Repeated questions than for New questions, $p < .001$, 95% CI [.22, .31]; and for Related questions than for New questions, $p < .001$, 95% CI [.18, .27]. No differences were found between Repeated and Related questions, $p = .54$, 95% CI [.00, .09].

There was also a main effect for question format, $F(1, 74) = 296.02$, $p < .001$, $\eta_p^2 = .80$, with students performing better on MC questions ($M_{MC} = .78$, $SD = .12$) than on SA ($M_{SA} = .58$, $SD = .17$) questions, $p < .001$, 95% CI [.17, .21]. Pairwise comparisons allowed to verify the same significant pattern of results ($M_{MC} > M_{SA}$) for each condition of type of questions (Repeated, Related, New), $p < .001$ (for all three conditions).

Finally there was an interaction between the variables type of question and question format, $F(2, 73) = 100.39$, $p < .001$, $\eta_p^2 = .58$.

To examine if student's performance on Repeated questions (on the Intermediate Test) is affected by question format on retrieval practice (i.e. Quiz), a one-way ANOVA for Repeated measures was conducted, $F(2, 73) = 32.49$, $p < .001$, $\eta_p^2 = .31$. There were no differences of student's performance for Repeated questions either they had been answered in MC or SA (on the Quiz), $p > .05$, 95% CI [-.0003, .09]. The difference of students' performance between New questions ($M_{new} = .61$, $SD = .13$) and the repeated questions that were answered in MC format on the Quiz ($M_{rep} = .75$, $SD = .21$) was greater, $p < .001$, 95% CI [.10, .18], d = .95, than the difference of students' performance between New questions and Repeated questions that were answered in SA format on the Quiz ($M = .70$, $SD = .19$), , $p < .001$, 95% CI [.05, .13], d = .65.

Congruency of question format

Repeated questions could appear on the Intermediate Test both on Same Format and on Different Format as the question on the Quiz. For example, after answering a multiple-choice question on the Quiz, students could answer it on the Intermediate Test on the Same format (MC) or on a Different format (SA). The same applies to Related questions, in which students could answer in the Same or Different format as their related questions on the Quiz.

To understand if students performed better on Repeated questions whose format was the same as on Quiz, a 2 (Question Format: Multiple-Choice, Short-Answer) x 2 (Format Congruency: Same, Different) repeated measures ANOVA was conducted. There was a main effect for question format, $F(1, 74) = 21.50$, $p < .001$, $\eta_p^2 = .23$, with students performing better on MC questions ($M_{MC} = .76$, $SD = .16$) than on SA questions ($M_{SA} = .69$, $SD = .22$), $p < .001$, 95% CI [.04,.11]. There was no effect of format congruency, $F(1, 74) = 1.09$, $p = .30$, $\eta_p^2 = .01$. However, there was an interaction effect, $F(1, 74) = 5.99$, $p = .02$, $\eta_p^2 = .08$. Pairwise comparisons showed that there was a significantly better performance when MC questions were on the Same format ($M_{same} = .79$, $SD = .21$) rather than on Different ($M_{dif} = .73$, $SD = .19$), $p < .05$, 95% CI [.02, .12]. No differences were found between students' performance on SA questions presented on Same format on Quizzes and Intermediate Test ($M_{same} = .67$, $SD = .24$) and Different format ($M_{dif} = .70$, $SD = .26$), $p = .36$, 95% CI [-.08, .03].

For Related questions, that could be answered in the Same or Different format as their related questions on the Quiz, the same analysis was conducted. To understand if students performed better on Related questions whose format was the same as their related question on Quiz, a 2 (Question Format: Multiple-Choice, Short-Answer) x 2 (Format Congruency: Same, Different) repeated measures ANOVA was conducted. There was a main effect for questions format, $F(1, 74) = 44.40$, $p < .001$, $\eta_p^2 = .38$, with students performing better on MC questions ($M_{MC} = .77$, $SD = .13$) than on SA questions ($M_{SA} = .64$, $SD = .18$), $p < .001$, 95% CI [.09, .16]. There was no effect for format congruency, $F(1, 74) = .66$, $p = .42$, $\eta_p^2 = .01$; nor there was interaction effect, $F(1, 74) = .30$, $p = .59$, $\eta_p^2 = .004$.

Accuracy for Repeated questions on Quiz and Intermediate Test

For Repeated questions, we examined whether students had previously (on Quiz) answered the questions incorrectly or correctly, as well as their performance on the Intermediate Test. Hence, students could have answered each Repeated question correctly twice ("double correct"), incorrectly twice ("double error"), or correctly once [answering correctly on Quiz and incorrectly on the Intermediate Test, getting worse ("decrement"); or incorrectly on Quiz and correctly on the Intermediate Test, getting better ("increment")]. We took also into account the fact that these questions could be on the Same or on a Different format.

On Table 3, it is observable that around 26% of the questions students had answered correctly on the Quiz, were also correctly answered on the Intermediate Test, both for Same format questions ($M_{same} = .26$) and for Different format questions ($M_{dif} = .26$). A 4 (Accuracy: Double correct, Double error, Increment, Decrement) x 2 (Congruency: Same, Different) ANOVA showed a main effect of accuracy, $F(3, 72) = 113.77$, $p < .001$, $\eta_p^2 = .61$, with "Double correct" answers being significantly greater than Double errors, $p < .001$, 95% CI [.13,.22], Increment, $p < .001$, 95% CI [.12,.19] and Decrement, $p < .001$, 95% CI [.16, .24]. Decrement was significantly lower than Double error, $p < .05$, 95% CI [-.05, -.01] and Increment, $p < .001$, 95% CI [-.07, -.02]. No effect of congruency effect nor interaction effect were found ($p > .05$).

Pairwise comparisons showed that the proportion of double errors was significantly larger for Same format than for Different format questions, $p < .05$, 95% CI [.00, .04], meaning that after answer incorrectly on the Quiz, students were more likely to answer also incorrectly if the format of the question is the Same on the Intermediate Test. Also, students

got worse ("decrement") more frequently for Different format questions than for Same format questions, $p < .05$, 95% CI [.01, .03], meaning that after answer correctly on the Quiz, students were more likely to answer incorrectly if the question format changed in the Intermediate Test. No differences were found for other Accuracy conditions regarding their congruency ($p > .05$).

Table 3 – Accuracy (proportion) for Repeated question in each condition according to format congruency

	Double correct	**Double error**	**Increment**	**Decrement**
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
Same format	.26 (.11)	.09 (.09)	.10 (.06)	.05 (.05)
Different format	.26 (.09)	.07 (.06)	.10 (.05)	.07 (.06)

Discussion

This study aimed to examine if the testing effect, as well as if the retrieval-induced facilitation was transferable to genuine educational settings. Testing effect was observable through higher students' performance for Repeated questions in comparison to New questions. Hence, results demonstrate the transferability of the testing effect for genuine educational settings, as previously established in literature (Bjork et al., 2014; McDaniel et al., 2013; McDaniel, Roediger, & McDermott, 2007; Glass et al., 2013). Moreover, results suggest a retrieval-induced facilitation of topically related information noticeable by higher students' performance for Related questions in comparison to New questions, as previously found (Bjork et al., 2014; Chan, McDermott, & Roediger, 2006; Chan, 2009, 2010). Contrasting with findings from Wooldridge et al. (2014), the present study suggests that students may benefit from retrieval practice for related non-tested topics at the same level as the Repeated questions themselves since no differences were found between students' performance on Repeated questions and Related questions. According to Wooldridge et al. (2014), teachers are not likely to repeat questions on a final exam identical to initial quiz questions. Based on our results finding, we can suggest teachers to use quizzes as retrieval practice and to use related questions on a final test, because students can benefit of testing even with related questions.

However, when we compared students' performance for Repeated, Related and New questions on the Intermediate Test, taking into account the questions format, we noted that the testing effect only occurred for SA questions. This may be, in part, due to the fact that generally MC questions, as a recognition task, are easier than SA questions (recall task), so no benefit is observed for MC questions, while for questions that are more difficult, it is beneficial if the topic has been already tested. Even if some studies find the testing effect with MC questions (e.g., Bjork et al., 2014; Kang et al., 2007), others found evidence that testing effect does not always occur when the final test is a recognition task (Chan & McDermott, 2007). More, Rowland (2014) found that recall final tests led to larger testing effects than recognition final tests, which is somewhat in line with the elaborative retrieval hypothesis (an effortful retrieval leads to the elaboration of the memory trace, making the information more likely to be retrieved later). Since the testing effect was more noticeable if the final task was a recall task, that is, a more effortful task, our results are congruent with this view.

SA questions on the retrieval practice (Quiz) led to a significant smaller testing effect than MC questions, with students in the Intermediate Test answering correctly 75% of Repeated questions that had been MC on quiz, against 70% of Repeated questions that had been SA, when compared to New questions (61% of correct answers). However, as Rowland (2014) stated, more effortful or difficult tests should, in fact, produce larger testing effects, if feedback is provided. In the present study, no feedback was provided, which may justify the results. Although feedback, either immediate (Butler & Roediger, 2008) either delayed by one day (Kornell, 2014) or one week (Mullet, Butler, Verdin, von Borries, & Marsh, 2014) is proven to nullify the interference of errors on later performance, enhancing performance, it is rarely given in educational settings. In the University where this study was conducted, teachers usually give feedback to students by meeting with them individually for them to see their errors, but students do not take this chance often. Considering this reality, by not providing feedback to students, this study portrayed what usually happens, thus, giving proof that even without feedback, testing students may indeed enhance their performance.

Besides, if the conditions of teaching and testing do not involve frequent feedback, multiple-choice quizzes may be the best option, as they may be considered a desirable difficulty in the classroom (Bjork et al., 2014).

On the Intermediate Test, students' performance was better when MC questions were previously answered on the Same format than if they had been SA questions before, but the inverse did not occur. This is partly congruent with the argument that MC questions might give students some kind of feedback (i.e. the four options) (Kang et al., 2007), since students

in the present study received no feedback after the Quizzes or the Intermediate Test. Also, on the Intermediate Test, MC questions still had the correct answer in the options to choose from, so students might easily recognize the correct answer. For the same reason, Related questions seem to benefit of the testing effect at the same degree, regardless of their format on the Intermediate Test. Applied to the testing effect, transfer appropriate processing hypothesis states that testing effect may mirror the cognitive processes utilized during the retrieval practice on a final test (in other words, if format on retrieval practice matches the format on the final test, it would be expected that matching tests would produce greater testing effects). Since students did not always benefit if the Repeated or Related questions on the Intermediate test were on the Same format as they were on the Quiz, these results do not seem to support the transfer appropriate processing explanation.

When we take into account students' performance on Quiz and Intermediate test, we understand they answer correctly both on Quiz and on the Intermediate test on around 52% of the Repeated questions, regardless of they being on the Same format or not (once more, results that do not support the explanation of the testing effect by the transfer appropriate processing hypothesis). Also, after answering incorrectly on Quiz, they get better on around 20% of the Repeated questions. More interesting is the fact that when they answered incorrectly on the Quiz, students had higher probability of making errors on the Intermediate Test if the questions were on the Same format. As students did not receive feedback after the Quiz, they might thought their answer was correct, keeping the same answer they gave previously. Likewise, as the Repeated question was in the same exact format might act as a cue for the students' previous (incorrect) answer. However, if the question format changed two things could have happened: 1) Even if students answered incorrectly on a SA question on quiz, the Repeated question would then be a MC question on the Intermediate Test, with four different options, one of them correct. 2) Students that answered incorrectly on a MC question on quiz, saw four possible options (one of them correct) and, when studying for the Intermediate Test, they could have used that information, answering correctly when the question was Repeated on the Intermediate Test, even on a SA format. The same rational may apply for the results showing that if they have answered correctly on the Quiz they had a higher probability of making errors if the question format changes: students could not be entirely secure of their answer on the Quiz, and when the questions were Repeated on a Different format, they would (incorrectly) change their answer, but they would maintain their answer if the question was Repeated on the Same format.

The testing effect occurred for SA questions regardless of the question format of retrieval practice. At the same time, Related questions also benefited from previous testing. These results are congruent with the idea that testing may promote the transfer of learning (Carpenter, 2012), either from one question format to another, either from tested information to related information.

As Glass et al. (2013) showed, students perform better in topics tested on a final exam even four to five months after the end of the course, boosting learning beyond the context and timeline of the college course. As long-term learning is the main goal of education, future research should focus on the long-term effects of testing in learning, through a final exam that evaluates all learned contents, for example.

Present results add to the growing research showing the benefits of testing beyond the laboratory and into educational settings. Benefits occurred not only for Repeated questions, but also for Related questions, and mostly for retrieval practice in the format of multiple-choice questions. This information may be useful to guide the way educators build evaluation methods in a way that promotes students' learning.

References

Agarwal, P. K., D'Antonio, L., Roediger, H. L., III, McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school student's test anxiety. *Journal of Applied Research in Memory and Cognition, 3*, 121-139. doi: 10.1016/j.jarmac.2014.07.002

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III, & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*, 861-876. doi: 10.1002/acp.1391

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2009). *Memory* (1st Edition). New York, NY: Psychology Press.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*, 201-210. doi: 10.3758/BF03193441

Bjork, E. L., Little, J. L., & Storm, B. C. (2014). Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory and Cognition, 3*, 165-170. doi: 10.1016/j.jarmac.2014.03.002

Butler, A. C., & Roediger, H. L., III (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*, 604-616. doi: 10.3758/MC.36.3.604

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*, 279-283. doi: 10.1177/0963721412452728

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*, 268-276. doi: 10.3758/BF03193405

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61*, 153-170. doi: 10.1016/j.jml.2009.04.004

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory, 18*, 49-57. doi: 10.1080/09658210903405737

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 431-437. doi: 10.1037/0278-7393.33.2.431

Chan, J. C. K., McDermott, K. B., & Roediger, H. L., III (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related

material. *Journal of Experimental Psychology: General, 135*, 553-571. doi: 10.1037/0096-3445.135.4.553

Gates, A.I. (1917). Recitation as a factor in memorizing. Retrieved from https://archive.org/details/recitationasafa00gategoog

Glass, A. L., Ingate, M., & Sinha, N. (2013). The effect of a final exam on long-term retention. *The Journal of General Psychology, 140*, 224-241, doi: 10.1080/00221309.2013.797379

Jacoby, Larry L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17*, 649-667. 10.1016/S0022-5371(78)90393-6

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III (2007). Test format and corrective feedback modify the effect of testing on longterm retention. *European Journal of Cognitive Psychology, 19, 528-558. doi:* 10.1080/09541440601056620

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62,* 227-239. doi:10.1016/j.jml.2009.11.010

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17*, 471-479. doi: 10.1080/09658210802647009

Kornell, N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 106-114. doi:10.1037/a0033699

Kromann, C. B., Bohnstedt, C., Jensen, M. L., & Ringsted, C. (2010). The testing effect on skills learning might last 6 months. *Advances in Health Sciences Education, 15*, 395-401. doi: 10.1007/s10459-009-9207-x

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morissette, N. (2007). Testing effect in the classroom. *European Journal of Cognitive Psychology, 19*, 494-513. doi: 10.1080/09541440701326154

McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*, 200-206. doi: 10.3758/BF03194052

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L., III (2013). Quizzing in middle-school science: Successful transfer performance on

classroom exams. *Applied Cognitive Psychology, 27,* 360-372. doi: 10/1002/acp.2914

Meyer, A. N. D., & Logan, J. M. (2013). Taking the testing effect beyond the college freshman: Benefits for lifelong learning. *Psychology and Aging, 28*, 142-147. doi: 10.1037/a0030890

Mullet, H. G., Butler, A. C., Verdin, B., von Borries, R., & Marsh, E. L. (2014). Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately. *Journal of Applied Research in Memory and Cognition, 3*, 222-229. doi: 10.1016/j.jarmac.2014.05.001

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review, 25*, 523-548. doi: 10.1007/s10648-013-9240-4

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*, 20-27. doi:10.1016/j.tics.2010.09.003

Roediger, H. L., III, & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249-255. doi: 10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., III, & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181-210. doi: 10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*, 1155-1159. doi: 10.1037/0278-7393.31.5.1155

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140,* 1432-1463. doi: 10.1037/a0037559

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 592.604. doi: 10.1037/0278-7393.4.6.592

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30,* 641–656. doi: 10.1037/h0063404

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6*, 175–184. doi: 10.1016/S0022-5371(67)80092-6

van den Broek, G. S. E., Takashima, A., Segers, E., Fernández, G., & Verhoeven, L. (2013). Neural correlates of testing effects in vocabulary learning. *NeuroImage, 78,* 94-102. doi: 10.1016/j.neuroimage.2013.03.071

Wheeler, M. A., & Roediger, H. L., III (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3,* 240-245. doi: 10.1111/j.1467-9280.1992.tb00036.x

Wing, E. A., Marsh, E. J., & Cabeza, R. (2013). Neural correlates of retrieval-based memory enhancement: An fMRI study of the testing effect. *Neuropsychologia, 51*, 2360-2370. doi: 10.1016/j.neuropsychologia.2013.04.004

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition, 3,* 214-221. doi: 10.1016/j.jarmac.2014.07.001

Yan, V. X., Thai, K., & Bjork, R. A. (2014). Habits and beliefs that guide self-regulated learning: Do they vary with mindset? *Journal of Applied Research in Memory and Cognition, 3,* 140-152. doi: 10.1016/j.jarmac.2014.04.003