Canadian Journal of Linguistics/Revue canadienne de linguistique (2024), **69**, 285–306
doi:10.1017/cnj.2024.24

CAMBRIDGE UNIVERSITY PRESS

THEMATIC ARTICLE

Improved student learning through active retrieval practice and random-sampled exams

Kaili Vesik and Kathleen Currie Hall

Department of Linguistics, UBC, Vancouver, Canada
Corresponding author: Kathleen Currie Hall; Email: kathleen.hall@ubc.ca

(Received 31 May 2022; revised 26 February 2024; accepted 6 March 2024)

Abstract
One pedagogical finding that has gained recent attention is the utility of active, effortful retrieval practice in effective learning. Essentially, humans learn best when they are asked to actively generate/recall knowledge for themselves, rather than receiving knowledge passively. In this paper, we (a) provide a framework for both practice and assessment within which students can organically develop active study habits, (b) share resources we have built to help implement such a framework in the linguistics classroom, and (c) provide some examples and evaluation of their success in the context of an introductory phonetics/phonology course.

Keywords: testing; learning; retrieval practice; software; randomization

Résumé
Une découverte pédagogique récente est l'utilité d'une pratique de récupération des connaissances qui est à la fois active et qui exige un certain effort. Essentiellement, l'être humain apprend mieux lorsqu'on lui demande de générer / rappeler activement des connaissances pour lui-même, plutôt que de les recevoir passivement. Dans cet article, nous fournissons un cadre dans lequel les personnes étudiantes peuvent développer, de manière organique, des habitudes d'étude actives; nous partageons les ressources que nous avons construites pour aider à mettre en œuvre un tel cadre dans une classe de linguistique; et nous présentons quelques exemples et une discussion de leur succès dans le cadre d'un cours d'introduction à la phonétique / phonologie.

Mots-clés: évaluation des connaissances; apprentissage; pratique de récupération d'information; logiciel; randomization

1. Introduction

We hope that a shared goal of educators is to help students learn material in ways that are both *durable* and *efficient* (Rawson and Dunlosky 2011). One finding that has gained recent attention is the utility of active, effortful retrieval practice in facilitating this type of learning (e.g., Brown et al. 2014, Rowland 2014). Essentially, humans tend

to learn best when they are asked to actively generate or recall knowledge for themselves, rather than receiving knowledge through passive techniques such as (re-)reading, highlighting, copying, etc. One interesting consequence of this is that the act of administering a test or an exam to students can itself help them learn the material—a phenomenon sometimes referred to as the "testing effect" (see Rowland 2014 for review and meta-analysis). As Roediger and Karpicke (2006: 181) explain, "testing not only measures knowledge, but also changes it, often greatly improving retention of the tested knowledge."

How, then, can instructors leverage this effect for better learning in their courses? In this paper, our goals are to (a) provide a framework for both practice and assessment within which students can organically develop active study habits, (b) share resources we have built to help implement such a framework in the linguistics classroom, and (c) provide some examples and evaluation of their success in the context of an introductory phonetics / phonology course.

Our approach combines several pre-existing pedagogical ideas into a novel form, facilitated by a purpose-built piece of software. We use sets of open-ended questions made available to students after each class session, which are then combined into individualized, highly customizable random-sampled exams by our open-source software. While many of the individual components of our approach have been used before, this software in particular is novel in that it allows for much more user-specified cross-categorization of topics and other randomization criteria (see section 4) than is typical of other similar open-source or proprietary applications. In turn, this allows for the other especially novel component of this approach, which is that the 'exams' in our courses are quite short (2–6 questions). The actual 'testing effect' component comes more from students' *preparation* for these randomized exams, where they are encouraged to test themselves using the full range of possible exam questions.

We present these ideas in the specific context of teaching linguistics. Although the strategies we describe are based on general principles of learning and should be applicable to all disciplines, we think there are a number of reasons why it is useful to describe them for linguists. First, linguistics is a field that lends itself to an open-ended, explanatory approach to learning because most 'real-life' applications of linguistics involve slower, analytical tasks rather than rapid exact recall. At the same time, this may lead instructors away from testing altogether, feeling project- or essay-based assessments to be more ecologically valid, and we think it's important to raise awareness of the utility of testing for learning. Also, it can be difficult to actively keep up with research in pedagogy in addition to one's own research specialty, and even if one does, it can be difficult to imagine re-shaping a linguistics classroom along new lines. Seeing concrete examples of how a system can work in a relevant context may make it more feasible to adapt to one's own needs. Finally, there is increasing awareness in the field of the need for more accessibility and inclusiveness, and there are aspects of this approach to learning that we think facilitate such goals.

The rest of this paper is structured as follows. Section 2 provides general background on the testing effect, and section 3 presents the basic architecture of our system. In section 4, we explain how we implemented the system with an open-access piece of software and how it was put into practice in an introductory phonetics and phonology course. In section 5, we look briefly at the effect of the system on

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

student grades and course evaluations. Finally, section 6 provides general discussion and conclusions.

2. The testing effect

The testing effect refers to a general phenomenon in which people are shown to have better long-term retention of material if part of their practice of studying that material involves some kind of effortful retrieval of the type often found on tests. It is important to distinguish 'testing-style practice' (which is what is important and the focus here; i.e., any practice that involves effortful retrieval) from 'actual tests' (i.e., examinations given in class and graded). While actual tests can give rise to the testing effect, the beneficial effect on learning can be seen with any testing-style practice, including self-administered testing. A canonical example is the use of flashcards for studying: flashcards typically require the user to actively recall specific ideas from memory, and hence facilitate learning more so than more passive approaches such as re-reading the same material.

There have been hundreds of studies on retrieval practice and the testing effect, too many to be reviewed here (see Agarwal et al. 2008 for a representative example). Rowland (2014) provides a meta-analysis of 159 such studies. 81% showed a positive benefit of testing-style practice, as compared to simple exposure, on subsequent recall. Other key findings of Rowland's analysis include: (1) feedback to students is crucial in order for the effect to emerge; and (2) the effect is stronger when (a) the material being learned is prose passages rather than single words and (b) the task involves recall (e.g., short answer) rather than recognition (e.g., multiple choice).

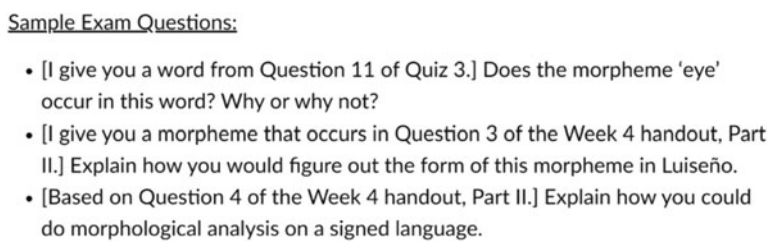
It's also important for instructors to realize that students both underestimate the utility of active testing practice and overestimate the effectiveness of passive approaches (e.g., Agarwal et al. 2008, Karpicke and Roediger 2008). Effortful retrieval is just that—effortful!—and if students do not understand that simple exposure will not achieve the same results, they are less likely to do it. As Roelle and Berthold (2017: 143) explain, "[t]he problem is … that learners scarcely engage in retrieval while performing learning tasks when it is not obligatory."

Taken together, these findings suggest that instructors who are teaching broad conceptual material (often presented in prose), as is common in linguistics, should actively encourage and facilitate effortful 'testing'-type studying and recognize the utility of in-class exams in student learning. This practice should include some kind of feedback to students, whether as conventional feedback from the instructor, or via techniques such as students checking their own work or having open-book tests (cf. Agarwal et al. 2008). Finally, Rowland (2014) shows that testing-style practice need not match the end goal to have a beneficial effect—so even if the end goal for a practicing linguist is to be familiar with general analytical concepts and how to apply them in a non-test situation, testing-style practice in the learning process is beneficial. This paper presents our approach to achieving these goals.

3. Basic architecture

Our approach involves regularly giving students specific short-answer questions that they know might appear on their graded exams, providing both motivation and structure for doing active, effortful testing practice. These questions are made available to

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press


Sample Exam Questions:

- [I give you a word from Question 11 of Quiz 3.] Does the morpheme 'eye' occur in this word? Why or why not?
- [I give you a morpheme that occurs in Question 3 of the Week 4 handout, Part II.] Explain how you would figure out the form of this morpheme in Luiseño.
- [Based on Question 4 of the Week 4 handout, Part II.] Explain how you could do morphological analysis on a signed language.

Figure 1. Excerpt from one day's sample exam questions.

students throughout the term, and students receive feedback by checking their own answers through reference to course materials or asking questions in class, tutorials, and office hours. The actual in-class exams are much less about comprehensively evaluating student knowledge and more about providing the enticement to engage in beneficial long-term study habits. The key features are that (1) exam questions are open-ended, (2) exams consist of a small, randomized subset of the material (with a *different* random selection for each student), and (3) exam questions are provided to students ahead of time, throughout the term. We cover each of these below.

3.1 Overview

We start by giving a brief description of what the exam structure in our classes actually looked like (see section 4.3 for more detail). After each class session (in our case, twice a week)[1], students were given a set of exam questions related to that day's material (see Figure 1; also discussed in section 3.2).[2] Students were encouraged to review the exam questions as they were posted, and could ask questions about them or their answers in class, tutorials, and office hours. They were never given a specific 'answer guide,' though they were given example answers to some questions (see discussion in section 3.2).

These questions were also added to a database of questions that would be used for exam generation, but was not itself made available to students. For each individual question, the database should contain all relevant information needed on a given exam. This includes content that students will see, such as instructions, data, and images. It also includes characteristics used to structure the random-sampling process, such as question source (e.g. 'Chapter 3'), topic, difficulty, and a unique ID number. Finally, each question's entry can contain instructor notes (e.g., answer keys) to be printed on instructor but not student copies. The database is presented to the script as a plaintext .tsv file but can be stored and edited in any format. For

[1]In our summer sections, each class session was three hours, and the course was six weeks long; in the fall, each class session was 1.5 hours, and the course was 12 weeks long.

[2]Note that these examples are presented exactly as they are given to students. That is, we do use a 'templatic' approach rather than listing out all of the specific questions that would have the same format. This is more efficient and flexible from our perspective and also increases the effort required by students to use the questions for studying. On an actual exam, a question like this might appear as: "Does the morpheme 'eye' occur in the following word? Why or why not? *spyglass*."

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

Topic	Source	Instructions	Data1_latex
Acoustics	Day 8 Handout, Question 7	Explain how each component of the description below gives you information about the specific sound being described.	This consonant typically starts off with nothing at all visible on the spectrogram. There is a short period of noise between the silence and the following vowel. This consonant typically brings down the second and third formants of the adjacent vowel.
Alternations	Quiz 7, Question 2	Explain why these statements about assimilation either are or are not true.	\begin{itemize} \item The process of assimilation is driven by articulatory factors, namely ease of articulation. \item Assimilation only applies to consonants. \end{itemize}
Phonological Features	Day 10 Discussion	Explain what the given feature's value is for this class of sounds, and why.	[approximant] / nasals

Figure 2. Excerpt of our question database for an introductory phonetics/phonology course.

example, ours was maintained in a shared online spreadsheet (see Figure 2) and exported to .tsv at generation time.

When it is time for an exam to be administered, a configuration file is created with the specifications for the exam. This is a plain text file that specifies the number of questions desired, the desired distribution of eligible topics (including 'wildcards') and difficulty levels, and the desired method of organizing questions. Users also provide a list of all the students for whom exams should be generated (e.g., using their student ID numbers).

The configuration file, student list, and question database are used as input files to our software, which then generates a unique randomized exam for each student, matching all the desired criteria. The software is a Python[3] script written specifically for this purpose, which is freely available under the GPLv3 license. The code itself as well as sample input files and more detailed information are available on GitHub.[4]

The software keeps track of what exams have been previously generated, so assuming that the size of the question bank permits, no student will see the same question appear more than once across all their exams in a course (e.g., quiz, midterm, and final). Furthermore, the questions for a particular student's exam can also be restricted based on those that appear on the exams of a particular group of peers (e.g., assignment partners). The exams can then be administered (and graded) in a variety of ways; see more in section 4.3.

[3]https://www.python.org/
[4]https://github.com/kvesik/examgeneration

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

The following sections dive a bit more deeply into the philosophy of each of the key components of this system, to highlight why and how each contributes to student learning.

3.2 Open-ended questions

The first key feature relates to the necessity for testing-style study to be effortful. Hence, the questions we provide involve an element of open-endedness that requires students to explain concepts in their own words (see Figure 1 for examples). Even in questions with clear-cut correct vs. incorrect answers, such as the first question in Figure 1, the second component of the question requires students to explain the rationale behind the answer, meaning they cannot just rely on memorizing specific answers to previously seen questions.

Open-ended questions are certainly not novel, though they are perhaps less common in large, introductory courses, where multiple-choice and similar questions are favoured. However, such closed-form questions do not really probe a student's understanding of the concept; a student simply has to be able to *recognize* the correct answer rather than *produce* it themselves (see e.g., Kang et al. 2007, Nedjat-Haiem and Cooke 2021).

Open-ended questions can be good for students in multiple ways.[5] First, of course, they encourage learning and long-term retention of material, as found in Rowland (2014) and described in section 2. As Nedjat-Haiem and Cooke (2021) discuss, there are multiple possible reasons for this. The act of retrieving the information itself seems to be beneficial, as shown by, e.g., Carrier and Pashler (1992), who showed that the same participants were better at recalling the second member of paired items when they were forced into a stimulus / response-retrieval mode during presentation than when they could just study both items simultaneously. Additionally, students seem to approach studying for open-ended questions differently (e.g., Thomas and Bain 1984, Scouller 1998, Martinez 1999, Struyven et al. 2005, Momsen et al. 2013). Students tend to use more 'deep'-style learning approaches, including various types of 'restructuring' original material such as comparing and contrasting ideas or coming up with additional examples, to prepare for more open-ended types of assessments. On the other hand, they use more 'surface'-style approaches, such as re-reading and memorization, which tend to reproduce the original material, for closed-ended assessments.

Note that a corollary of this tendency is that there is utility in thinking about what type of end understanding or abilities one wants students to attain, and then designing assessments to match that goal. If students "learn the forms of knowledge and develop the cognitive abilities that they are asked to demonstrate [in assessments]" (Scouller 1998: 454), then our choice of assessment has longer-term consequences beyond evaluation of their immediate mastery of course material. If as linguists, we want students who can evaluate evidence, formulate argumentation, and connect various ideas together, then those are the skills we should ask them to demonstrate

[5]While there is still a beneficial testing effect for multiple-choice tests, exposure to incorrect choices on a multiple-choice test can also impair longer-term understanding; see, e.g., Roediger and Marsh 2005.

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

in our courses. This is different, for example, from a profession in which an ability to quickly recall factual information would be especially beneficial, for which different kinds of training and assessment would be more useful (e.g., at least certain aspects of the medical field, where quick recall of anatomy or symptoms associated with particular conditions is necessary).

Second, open-ended retrieval questions also maximize the student's ability to take ownership of and get credit for their understanding (e.g., Hubbard et al. 2017), which we see as an important aspect of 'optimizing relevance,' part of the guidelines for a "Universal Design for Learning" (Meyer et al. 2014). For example, Figure 3 contains five different, but all satisfactory, answers to the same question, allowing students to demonstrate additional or alternative understanding beyond simple transcription. Note that the answers incorporate all sorts of *different* types of information, from articulatory phonetics to allophonic processes to dialect variation. By not forcing students to produce one 'right' answer, the instructor lessens the chances of penalizing students whose thought processes are different from their own.[6]

We also think it is important to provide concrete examples of answers to these kinds of questions (as in Figure 3, which shows some examples we provided to our students). Providing exemplars not only illustrates the depth of answer being sought but also exemplifies that radically different answers can be considered correct. This is an idea that some students have expressed discomfort with, wanting to know more concretely what is 'expected' of them. Providing specific examples of very different answers to the same question seemed to help our students feel more confident that their own interpretation and understanding was valid. While we did not specifically build in time to go over such sample answers during class, such engagement would likely add to their utility. In a meta-review of the use of this kind of example, To et al. (2022) find that the best results in terms of student confidence and academic performance come when there is active engagement with the exemplars on the part of the students. In particular, they recommend a method whereby students first produce their own answers to a sample question and then engage with the example answers, as this can help ensure that students are still taking ownership of the answers and encouraged to think creatively, rather than feeling they need to conform to the content of the exemplars.

3.3 Randomized, subset exams

A typical downside to having open-ended questions requiring student explanations is that they can be time-consuming to grade (Martinez 1999, Haudek et al. 2017, Hubbard et al. 2017). To counteract the loss of efficiency in having open-ended questions, the use of randomized exams is key. Each exam consists of a small number of questions (the maximum we've used is six), with each question covering a different topic, and the questions being a different random selection for each student. For instance, of the sample questions on morphology shown in Figure 1 (along with any other questions on morphology throughout the term), no student would get

[6]At the same time, open-ended questions like this are admittedly better at evaluating which aspects of a concept a student correctly understands, rather than diagnosing common incorrect understandings (see discussion in Hubbard et al. 2017).

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

Question 1: Transcription
Source: Week 2 Handout, Part II, Question 11

How would this word be transcribed? Explain how you chose the last symbol in the word.

<little>

****

Possible Answer #1: [lɪɾl̩]

The last sound is transcribed as a syllabic [l̩]. I picked this symbol because I don't think I have a full vowel in the second syllable of the word; I think the [l] sound acts as the heart of the syllable.

Possible Answer #2: [lɪɾl̩]

I used [l̩] as the last symbol in this word to represent the syllabic [l] sound. This word is similar to words like <puddle> and <middle>, which have syllabic [l̩] at the end.

Possible Answer #3: [lɪɾəl]

I used [l] as the last symbol of this word. It is the same as the first sound in the word, and both are voiced alveolar lateral liquids, for which the IPA symbol is [l]. While there might be some difference in the details of how the sounds are pronounced in these two positions, I think they are both captured by that articulatory description, and we generally use broad transcriptions in this course.

Possible Answer #4: [lɪɾəɬ]

I used [ɬ] as the last symbol of this word. While it is similar to [l], I know that this sound is usually pronounced as a "light" [l] sound in initial position but as a "dark" [ɬ] sound in coda position / after a vowel.

Possible Answer #5: [lɪtəl]

Note that my dialect is not Standard North American English; I usually do pronounce a [t] in the middle of a word like this instead of [ɾ]. In terms of the last symbol, I picked [l] because even though I know it could be syllabic, I think that using a [t] sound makes the vowel clearer, so I think that there's a vowel followed by a regular [l].

Figure 3. Example possible answers (shared with students) to an open-ended question about phonetic transcription. In our class, these answers were all evaluated as equally good, but of course other instructors might ask a more specific question and then value various answers differently.

more than one on an exam. The student has no way of knowing which question they will get ahead of time, so they get the benefit of *studying* all the questions. However, the grader(s) have to evaluate only one answer per topic per student, streamlining grading (as compared to having to grade multiple questions on each topic for each

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

student).[7] And, another component of efficiency is the development of the questions in the first place—it can be faster to come up with a set of open-ended questions than to devise multiple-choice questions, where both the questions *and* the answers have to be developed ahead of time.

The questions included will inevitably be different in their scope and subjective difficulty level, both inherently and because of things like whether specific answers have been discussed with students. Our software allows more flexibility than typical randomization tools. As described in section 3.1, each question can be tagged with topic, difficulty level, and other key information. In turn, the question generator can create an exam that has, for example, one question on phonetics and one on morphology, and, orthogonally, one easy and one hard question, and, orthogonally again, no more than one question from a given language. With a sufficiently large database, all these parameters can be met, such that exams should be roughly comparable in terms of 'fairness' across students, despite randomization (see also section 5.3).

3.4 Question availability

One potential concern with using random sampling in the way described above is that, if we begin with the premise that testing itself enhances learning, then using such a small number of selected questions on each exam could mean that only those few, specific topics will be 'learned' in the long term. This leads to the third key feature of our approach: while students cannot know ahead of time *which* questions they will be asked, they have access to the database of exam questions in advance. Crucially, the database is not simply provided *en masse* shortly before the exam, but rather, new, relevant questions are provided after each class session with the intention that students will use those questions consistently throughout the term in order to *test themselves* and thus take advantage of the testing effect beyond the utility of the in-class exams. And, in our classes, they seem to do exactly this; at least one student showed up to office hours with an entire list of the questions and their own sample answers, and others certainly asked questions suggesting they were doing something similar (see also discussion in section 4.3).

While there is no magic formula for forcing students to engage in such behaviour, providing questions on a regular basis gives students both (a) a concrete study method that encourages them to test themselves and get help if they need it and (b) materials to use for studying, thereby increasing their chances for successful learning. That said, we do think it would be even more beneficial to provide students with more explicit guidance on the theory behind this approach to exams and to give them chances to practice studying and getting feedback on their responses in tutorials; this is not something we specifically did in the courses described here, but is how we are now implementing this approach.

Note that there is an important difference between providing students with the actual exam questions and giving them a list of 'key concepts' to study. While a

[7]Having different questions for each student may make things somewhat slower than having a stack of identical questions. But, with a very small number of questions per student, and similar *types* of questions across students, our experience is that grading is faster overall than with our traditional exams.

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

list of topics may focus students' attention on the right ideas, it does not provide any additional support for *how* to study. A student could, for example, re-read the relevant textbook sections or re-watch part of a class lecture, instead of retrieving the knowledge for themselves. By framing the items in terms of questions, we give students a framework in which to actively try to answer those questions as part of their study habits and hence benefit from the testing effect.

There are additional advantages to providing questions ahead of time. There are many different ways that exams can be structured, and students come in with different degrees of experience with a variety of question types. Those with wider experience may have an advantage when it comes to performance on an exam, simply because of their ability to anticipate likely exam questions, regardless of any actual difference in mastery of the material. By providing students with the actual questions that will be asked, we put them on a level playing field in this regard.[8] In addition to our own observations along this line, another professor in our department, who was in the process of implementing this approach for the first time, reported to us that a (graduate!) student specifically commented on the utility of the review questions to create a guided study experience, as compared to passively reviewing material. Furthermore, providing questions ahead of time reduces the 'unknown' aspects of an exam that can contribute to test anxiety. These benefits, combined with the supply of material to encourage self-testing/retrieval practice, means that the structure of the assessment style in the course should facilitate *everyone's* long-term success.

We should stress, too, that providing students with the questions in advance does not mean that there is no test of a student's ability to apply concepts to new material—this is just done outside of exams themselves. For example, in our course, there is a weekly handout of problems that guide in-class discussion. It is rarely the case that all parts of all questions on the handout are covered in class, but exam questions can directly reference those un-discussed items. This gives students motivation to try out the additional exercises and makes them more likely to seek help or feedback at an appropriate time. Once we started explicitly listing questions about these exercises as possible exam questions, we had more students coming to office hours and tutorials having specifically tried these questions on their own and asking us for additional feedback than we had ever had before, even though we have always stressed the usefulness of trying these questions for practice. Additionally, for the final exam, we provided a new dataset and possible accompanying questions—students were told that they could discuss it amongst themselves and ask general questions, but no feedback was provided about the correct analysis (or even what elements were to be analysed). This question allowed us to push students to apply knowledge in new ways, and seemed to be effective. It was very clear from questions we got before the exam that many students were indeed working through the problem (e.g., they would ask about a general theoretical concept that would be helpful in analysing the dataset, even if they avoided asking about the dataset itself), and many were able to

[8]Again, this is not an entirely novel practice, but we are more familiar with it in the context of instructors giving students a list of essay questions, some number of which students should expect to see on an exam. What's different here is that we are providing all questions, regardless of type, ahead of time and on a daily basis.

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

successfully work through a complete analysis on a previously unseen dataset. At the same time, it was also effective at identifying students who really did not understand the underlying concepts, as they tended to completely fail these questions on the exam. Of course, we'd *rather* have all students succeed, but the failure of some students is a useful barometer of the principle; giving students questions ahead of time does not just automatically lead to across-the-board perfect grades—students still need to put in the work to answer them.

To summarize: we have found that providing students daily with specific, open-ended questions that will form the basis for a randomly-sampled exam actively encourages good study habits in the form of effortful retrieval practice along with feedback. This approach allows instructors to target skills and concepts relevant to linguistics and allows students the opportunity to take ownership of their learning and be on a more level playing field when it comes to approaching exams.

4. Implementation

In this section, we go into a bit more detail about the question bank and Python script, and then demonstrate more specifically how we have used this approach in an introductory phonetics and phonology course.

4.1 Question database

The question database used by the exam generation script is built by the user to custom-suit the course, following the template provided with the script. The development of the database is by far the most time-consuming—and most important—part of this assessment approach; however, the work involved can be distributed over the term. Not only do the questions need to be thought-provoking and numerous enough to motivate students to practice consistently throughout the course, but they also need to cover a wide range of values across several dimensions for the randomized nature of the exams to be fully feasible. For instance, if exams are to include questions from five different topics across three different difficulty levels, then each of those fifteen combinations should have at least one question—preferably several—to draw from. (For reference, our database included about 400 questions, though many were 'versions' of the same question—e.g., sixteen of them asked the same question about how to transcribe a word, each focusing on a different word.)

4.2 Python script

The exam generation script provides a text-based interface that is run from the command line or Python console. The user inputs the path to a plaintext configuration file (created by the user) with parameters for the construction of a session's exams (as described in section 3.1). The software then outputs student copies, instructor copies, and a question bank as LaTeX source (.tex) files that the user can compile into .pdf format (see, e.g., Figure 4), using a regular LaTeX editor. No knowledge of Python programming is required to interact with the script, but one does have to have Python installed and know how to open and run a script. LaTeX can be either

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

Question 2

Source: Day 6 Handout, Question 11

What do the two images below tall you about the phonological status of handshape in ASL, and why?

(a) STAY

(b) AWKWARD

INSTRUCTOR NOTES: Nothing, because both handshape and movement are different.

Figure 4. Sample layout of a question from a generated .pdf exam. (ASL images from ASL-Lex 2.0, Sehyr et al. 2021). Note that the student copy would be identical to the above, except without the "Instructor Notes" section.

installed on the user's local machine or used online, and use of our software requires minimal to no knowledge of LaTeX markdown.[9] We include documentation on the project GitHub site for how to set up Python, run the script, and use LaTeX to generate .pdfs, including the minimal knowledge one needs to be aware of to ensure basic formatting.

4.3 Example usage

One advantage to this system is that it is customizable for a wide variety of situations and uses. In the three sections of the same course we have used it in so far, we have tried seven different implementations that fall into three broad categories: (1) a written exam, (2) an oral exam, and (3) an oral 'quiz.' Below, we do give the very specific details of how we used the strategy, but our intention is that instructors shape the implementation to their own needs, while incorporating the particular elements that are in keeping with the larger philosophy of encouraging student learning via effortful retrieval.

In all cases described here, assessments were administered as part of a fully online, introductory phonetics and phonology course (LING 200) at the University of British Columbia-Vancouver (UBC), during the COVID-19 pandemic. The course had always previously been taught in person, and very little about the delivery changed during the online conversion other than the assessments themselves. Enrollment in the course ranged from N = 42 in the summer to N = 136 in the fall.

[9]Note that a user *may* use markdown to format questions and formulae, etc., but does not have to. For instance, the itemized list in the 'Data1_latex' column of Figure 2 above could have been input as a screenshot from a separately formatted source, instead of formatted in the database using LaTeX markdown.

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

Exams consisted of 5–6 questions per student; quizzes had only two. All were explicitly 'open book.' Students were informed in advance about the topic and difficulty distribution, in addition to being given access to possible questions throughout the term as described in Section 3.4. We included constraints against the repetition of topics and questions within and across exams.

For written exams, the software generated a separate document for each student, with one question per page, in a pre-set order; each student got access to their own question document online. All questions were required to be answered during a 45-minute window. For the oral exams and quizzes, students signed up for an online time slot (15 minutes for exams, 5 minutes for quizzes) with the instructor, during a multi-day window. In this case, we had the software generate a single document containing all exams for a given day's schedule, with one question per page, in a random order within each exam. During the assessment itself, the instructor shared the .pdf containing the questions on screen and advanced through them one at a time. If students made a concerted effort to respond to each question and move forward, there was no obligation to complete the exam; instead, grades were based on the questions actually answered. The instructor did not provide overt feedback, but did ask clarification and follow-up questions as warranted to better gauge the student's level of understanding.

We found that grading this style of exam was surprisingly easy, despite the lessening of 'flow' that one often achieves when grading a stack of identical questions. We decided to assign all questions the same value,[10] regardless of difficulty, to help even out potential inequalities. Specifically, we were worried that these might arise due to the random selection of questions or subjectivity in decisions about 'difficulty,' as in our case, difficulty levels were based on our own impressions from experience teaching this material over the years, and not, e.g., determined by average past performance on the same questions or student ratings. In our classes, we had only the primary instructor grade all of the questions on all exams, to mitigate variability (especially with oral exams). The ease of grading depends a bit on the format of presentation. For oral exams, grading can be done essentially in real time, with a bit of time after the fact needed to actually enter grades and comments. For written exams, grading is easiest when students can write in their answers immediately below their own questions, rather than having a 'template' that all students use, because the latter approach necessitates cross-referencing the original questions. Although there is a subjective component to grading open-ended questions, we received no complaints about the specific grades students earned on any of the seven exams across three terms; see also section 5.3 for discussion of student perception of fairness.

Choosing between oral vs. written exams and open- vs. closed-book exams are topics beyond the scope of this paper, and have been discussed extensively elsewhere (e.g., Huxham et al. 2012, Iannone et al. 2020, and Theobold 2021 on oral exams; Theophilides and Dionysiou 1996, Agarwal et al. 2008, and Roelle and Berthold 2017 on open-book exams). Note that it is not yet clear whether there is a learning advantage to having either open- or closed-book exams; e.g., Agarwal et al. (2008)

[10]Details of the rubric used are available on the project's GitHub site.

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

found no particular longer-term advantage for either exam type. Given (1) the reported psychological benefits to students (e.g. Theophilides and Dionysiou 1996), (2) the fact that being explicitly open book reduces risk of overt examination misconduct, and (3) the fact that most 'real-world' linguistics applications allow access to resources, we personally prefer the open-book approach. An open-book exam does itself provide feedback to students in a way that is thought to induce the testing effect for the material actually on the exam (Rowland 2014). Another advantage to our approach is that it is relatively simple to switch between oral and written exams, in either direction (see also Waterfield and West 2006, section 5.5, for a more careful examination of a preference for oral rather than written exams among students with disabilities, especially dyslexia). Impressionistically, while many students preferred the familiarity of written exams, they generally performed better on the oral exams (the variability in implementation means that statistical comparison isn't appropriate, but the median scores on oral exams ranged from 5 to 10% higher than those on written exams).[11] At the same time, the scores on both written and oral versions ranged widely, from failing grades of 20–30% to excellent marks of 100%, suggesting that the randomized style still allows for differentiation among students.

Furthermore, this approach seemed to incentivize exactly the behaviour we want from students. From our impressions as instructors, they engaged with the material and seemed to consistently review it on their own, practicing new applications, asking about it in tutorials, and discussing it with each other. It felt as though students were empowered to take control of their learning because we had given them a structure that naturally supported best practices in terms of study habits. For us, the change was especially noticeable in the kinds of questions students were asking. We got fewer questions about minor points within the reading, what would be on an exam, or what the "final" answer of a question might be, etc. We also got more questions that showed students were re-engaging with previous problems and trying to better understand the underlying principles behind them, and attempting the additional practice problems on their own (these questions sometimes were generic questions such as "Can we go over problem X?" and sometimes more explicit questions such as "Would this be a good explanation for how I came up with the underlying form?"). In large part, this is because we literally told them which questions would be on the exam: e.g., they knew they might have to explain how to figure out the underlying representation of a particular morpheme in a dataset, and would not have to explain a minor concept that was never discussed in class. But telling students what the questions will be is not the same as giving them the answers; it is just being more explicit about what components of a practice problem or class discussion they should be working to understand. Students are then able to direct their energy to the right areas.

5. Quantitative and qualitative analysis

There are many caveats to doing statistical analysis on observational course data, especially data generated during such an abnormal time as the years 2018–2021,

[11]It is beyond the scope of this paper to examine what to attribute this difference to, but student studying behaviour, the interaction between student and instructor on the exam itself, and instructor grading behaviour are all possibilities.

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

including the beginning of the COVID-19 pandemic. Observational data is not controlled experimental data—the students and instructional teams in the courses are different; the method of delivery (face-to-face vs. online) varies; the external stresses on everyone differ radically; the implementation details of exams and course content vary; etc. That said, it is useful to see whether there are any obvious trends that would indicate either an advantage or a disadvantage for students who experienced the randomized approach to exams as compared to a 'traditional' (non-randomized) approach. The 'traditional' exams here consisted of both open- and closed-ended questions; they were closed-book; and they were written in person during class time. Most crucially, they did not consist of a small random sample of questions unique to each student; all students received the same exam and did not know the possible questions ahead of time.

We examined two grade-based outcomes as well as student evaluations of the course. The two grade-based outcomes were the effect of traditional vs. randomized exams on (1) the final course grade within the LING 200 course itself, and on (2) exam and final grades in LING 311, a subsequent upper-year course focusing on phonological analysis, for which LING 200 is a prerequisite. We then also examined student responses to course evaluations in the traditional and randomized versions of LING 200. This retrospective study was carried out under approval from the UBC Behavioural Research Ethics Board (#H21-02730).

We attempted to control for as many variables as possible. This included examining data from only summer term sections of LING 200, to maximize consistency of enrollment and pace of course. All sections, both with traditional exams (summer 2018 and 2019, in-person) and randomized exams (summer 2020 and 2021, online) were taught by the same instructor (author KCH), and the rest of the course content was as equivalent across the four sections as such things can ever be. Grade-based data were included from all students in the courses, except three students who earned zeroes on the final exam.[12]

5.1 LING 200 results

Figure 5 shows final course grades for students in LING 200, grouped by course format (traditional vs. randomized exams). Overall, there is no obvious difference in the final course grades based on exam type, and this was confirmed with a simple t-test comparing final course grades across the two types of course [$t(131.14) = 0.39$, $p = 0.70$]. To the extent that grades reflect knowledge, then, there seems to be no immediate benefit or detriment to the use of randomized exams. Given the number of factors working against students in the terms where randomized exams were used, it is at least encouraging to see that there was no serious negative effect of this radical change, though of course it would have been even better to see specific increases (see also e.g. Supriya et al. 2021 for related discussion of student perceptions and academic performance during COVID).

[12]The gathered data did not provide information about whether these students (one from a traditional section and two from a randomized section) simply failed to take the final exam or actively failed to earn any points in their attempt.

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

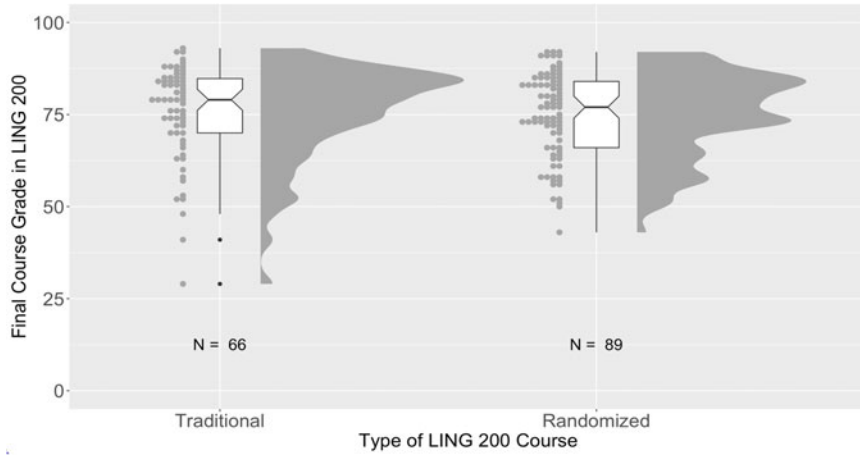


Figure 5. Distributions of final LING 200 course grades for two years' worth of traditional exams and two years' worth of randomized exams.

5.2 LING 311 results

We next consider the grades from LING 311. Again, to maximize control, we collected data from only the sections of LING 311 that were offered in the fall terms of 2018, 2019, 2020, and 2021. Three of these sections were taught by a single instructor; one term, fall 2019, was taught by a different instructor. Neither of the two instructors was involved in the delivery of any section of LING 200 considered here, and both were full faculty members with experience teaching both courses at UBC. All four sections had traditional-style exams; the 2020 section was online, and the exams in 2020 and 2021 were open book.[13]

Data were gathered from all students in the courses, but some data points were subsequently excluded. First, as above, data for students who earned zeroes on the final exam were removed (N = 3). Second, data for fall 2021 students were excluded if we did not also have their data from LING 200, as we had no way of determining whether they experienced a traditional or randomized LING 200 section (N = 36).[14] Finally, we included only students who took the two courses in immediately consecutive terms. Although it removed a large number of students from the analysis (N = 147), we felt that the confounds related to this factor (e.g., the variable length of time between courses, the different attitudes of students who choose to take the courses back to back vs. separating them, etc.) outweighed the benefits of larger numbers.[15]

[13]We acknowledge that there is still a large amount of variability here that impedes our ability to get a clear look at the effect of LING 200 type. However, as we think that the effect on future courses is a better measure of long-term learning than changes within the course, we think it is still worth making the comparison.

[14]Note that the same was not true for students in fall of 2020; if they did not take LING 200 in the immediately preceding summer term, then they must have been in a traditional-style LING 200 course, even if we did not have their actual data from LING 200.

[15]Note that the trends if all students were included are similar to those shown in Figure 6, and in neither case are any differences statistically significant.

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

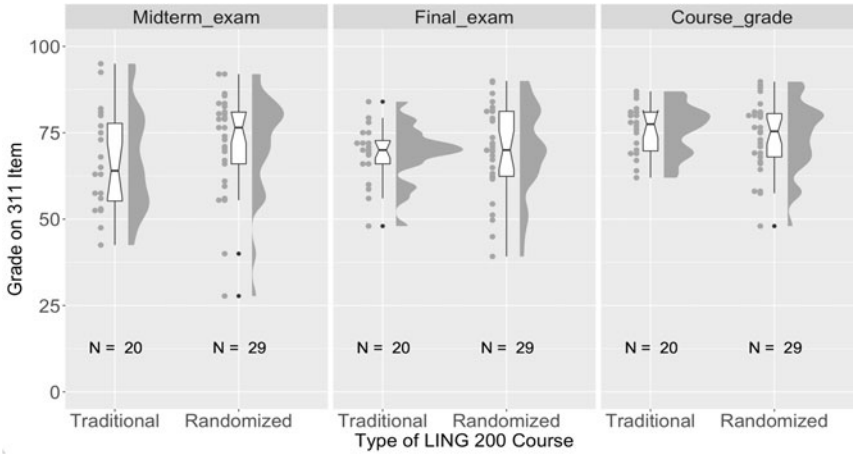


Figure 6. Distributions of midterm exam, final exam, and final course grades in LING 311 for two years' worth of traditional exams and two years' worth of randomized exams in LING 200.

Figure 6 shows the results on three different components of LING 311: the midterm exam, the final exam, and the overall course grade, separated by the type of exam experience students had in LING 200. Among these students, there seems to be a small benefit of randomized exams in LING 200 on the midterm exam in LING 311, but no difference post-midterm, although the effect of randomization does not reach statistical significance. This was confirmed using a linear mixed-effects regression model predicting grade from type of LING 200 course and item in LING 311, with random intercepts for each student. This model was compared to one without the LING 200 course type being a predictor, and the two were not found to be significantly different assuming an α of 0.05 [$\chi^2(3) = 6.69$, $p = 0.08$].[16] The fact that the effect (such as it is) is limited to the midterm exam suggests that it might have more to do with better recall (i.e., enhanced learning) of the earlier material in the course, which is more directly related to material in LING 200 than materials later on, than it does with any longer-term change in study habits of the students.

5.3 Student evaluations

We additionally examined student responses to the course evaluations for the randomized LING 200 courses. In retrospect, it would have been even better to have a customized questionnaire that specifically addressed the exam structure, to get a better picture of how student behaviour may have changed, but in the absence of this, we can use the standardized questions as a proxy. Again, we focus on the summer sections described above. However, not all students filled out the questionnaire; of the 69 students in the traditional sections in 2018–2019, 34 students (49%) submitted responses, and of the 91 students in the randomized sections in 2020–2021, only 30 students (33%) did. Hence, the following data may not be

[16]See the supplementary materials for details of the two individual models.

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

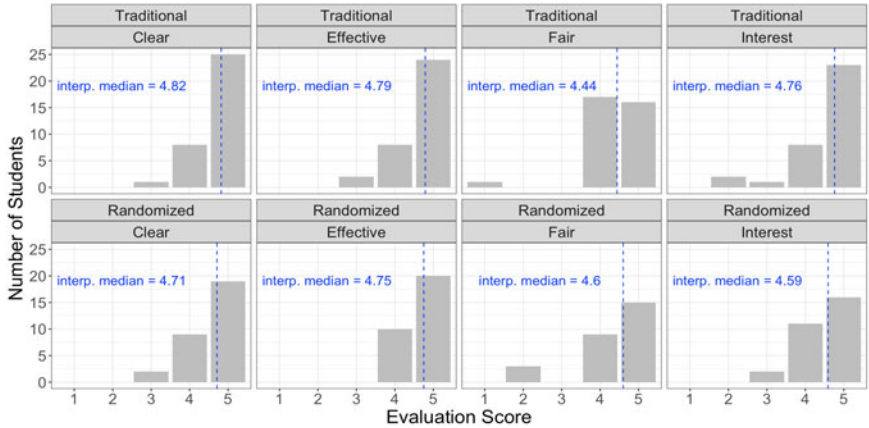


Figure 7. Distribution of responses and interpolated medians for four questions on student evaluations of teaching, for two years of traditional exams (top) and two years of randomized exams (bottom). See main text for details of the four questions asked.

truly representative of student impressions. Additionally, we note that "[s]tudents typically are not well situated to evaluate pedagogy" (Stark and Freishtat 2014: 6) and that there are many concerns with trying to interpret student evaluations of pedagogical 'effectiveness,' so these responses should be interpreted as reflecting student feeling rather than, e.g., learning outcomes.

We focused on (1) the answers to four specific prompts posed by the university that seemed most germane to student experiences of the difference between traditional and randomized exams and (2) the free-form responses that students gave that mentioned anything about aspects of the course that differed between the two course types. The four specific prompts examined here were (with key words used in Figure 7 bolded): (a) "The instructor made it **clear** what students were expected to learn"; (b) "The instructor communicated the subject matter **effectively**"; (c) "Overall, evaluation of student learning (through exams, essays, presentations, etc.) was **fair**"; and (d) "The instructor helped inspire **interest** in learning the subject matter." Students could respond to these using a five-point scale, with '1' meaning 'strongly disagree' and '5' meaning 'strongly agree.' Following the guidelines at UBC, we use the interpolated medians as the most useful summary measure of these scores; the interpolated median takes the frequency distribution into account, which we also visually illustrate in Figure 7.

As can be seen in Figure 7, both the distributions and the interpolated medians for the two types of courses were quite similar for each of the four questions. A linear regression predicting score from prompt and course type showed that there were no significant differences; the responses to the four different prompts were not different from each other, the responses from the two course types were not different from each other, and the interactions between the two were not significant.[17] The one element that was close to being significant was that the scores for the question on 'fairness' were slightly lower than the

[17]See the supplementary materials for details.

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

scores for the question on 'clarity' ($t(7, 244) = 1.93$, $p = 0.05$; note that 'clarity' was simply our arbitrary baseline question), but interestingly, the students in the *randomized* sections had a smaller difference than those in the traditional sections (though again, this difference was not significant; $t(7,244) = 0.363$, $p = 0.72$). This is also the only prompt for which there is also a difference in the distributions; under the traditional approach, there were more '4' responses than '5's, while in the randomized approach, there were more '5' responses than '4's. Together, this suggests that if anything, students found the randomized approach to exams a *fairer* examination structure than the traditional approach.

Finally, we can look at the free-response comments from students. While we did specifically ask students to address the format of exams, very few students did so, and when they did, they more often commented on the use of oral as compared to written exams rather than the use of randomized exams, so we have few comments to report here. This suggests that students were not particularly affected either positively or negatively by the switch. We include some comments below to give a sense of both the positive and negative concerns from students; note that these also include comments from the fall section of the course, unlike the scores shown in Figure 7.

- "My least favourite part of the course was the oral mid–term exam. However, I appreciate how [author KCH] went above–and–beyond to ensure that it was fair."
- "I was nervous about the oral midterm, but I do agree that it was a fair way of assessment. I liked how each topic had a handout sheet with lots of practice questions. I understand why Prof Hall doesn't have answer sheets, but for the trickier questions, I think it would be helpful to have an answer sheet posted after a certain amount of time. I liked how she included student examples of a good analysis."[18]
- "The post class emails and sample exam questions were really helpful to summarise the content and keep caught up with all aspects of the course."
- "Everything was doable and really tested your knowledge of course content. It was nice that all the content connected, by the end of it I found myself using aspects I learned in week 1 in a week 11 assignment."

6. Discussion and conclusions

Our aim in this paper has been to summarize some of the literature on the testing effect, i.e., the utility of effortful retrieval practice in improving learning, and to explain one approach to facilitating such practice in the linguistics classroom. Our system involves providing students with open-ended exam questions after every class session, with the intent that students use these questions to structure their study sessions and test themselves. We use custom-built software to create short, random-sampled exams that balance subject, difficulty, and other factors, which allows flexibility in exam administration and grading. Although we have not

[18]Note that although several students requested answer guides, providing them would seriously compromise exam integrity given the randomized structure and open-book policy on exams. As this student noted, we did provide *examples* of good answers instead.

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

conducted a controlled experiment based on this technique, there do not seem to be any obvious negative consequences to the approach, either in the introductory course where it was implemented or in the subsequent higher-level course. Impressionistically, it seems that students are indeed focused on the type of engagement, studying, and learning that we would like to see in our classrooms.

In addition to the strict learning benefits, we also think this approach increases accessibility and intersectionality in the classroom. The open-ended nature of the questions allows students to draw on their own experiences and understanding and highlight how they think about the topics, rather than forcing them to think about material in exactly the same way as the instructor. The fact that the actual exam questions are available on a daily basis removes some of the anxiety around unknown exam content and structure, and gives all students access to optimal study strategies.

We think that these techniques could be useful in any classroom, but we have highlighted their utility for linguistics: the type of subject matter taught in linguistics classes is likely prone to the benefits of the testing effect, and even though 'linguistics in the real world' does not necessarily involve test-like situations, the act of testing the type of knowledge we want linguists to have is likely to increase students' learning of it.

We do think that it would be beneficial to probe the benefits of this approach more systematically; we implemented it during the COVID-19 pandemic, largely driven by necessity rather than careful design. It would be useful, for example, to evaluate the effects beyond numerical achievement scores and get direct feedback from students about their personal experiences in this system. Are they in fact using the questions for self-testing as intended? How often and in what contexts do they look at them and practice with them? Do students typically write out answers and then refer to them on exams, or simply practice the material orally / mentally? Do they feel that having the database of questions helps them understand the expectations in the course? Do they feel their engagement is the same or different than it would be with a traditional exam? Does the fact that they know the exam will have just one question per topic mean they skipped over some topics in studying? These are all questions that it would be useful to have student answers to rather than our anecdotal impressions.

In conjunction with this information seeking, we think it would be useful to instruct students more directly about the pedagogical aims of the approach and how they can best make use of the resources provided, especially given the evidence that students are often not very good judges of study-practice effectiveness. In our classes, we mostly just presented this approach as a *fait accompli* rather than spending time discussing the intentions and benefits, but in retrospect, it would probably be beneficial to students to have more concrete information about our intentions. Rawson and Dunlosky (2011) provide a very specific recommendation for how students should go about self-testing: they suggest that it's best to aim for the ability to correctly answer a question three times during the first learning session, and then to have three subsequent "re-learning" sessions in which the question is answered correctly once per session. We have not made any recommendations of this sort to our students—just provided them with the questions—but it would likely be helpful to them to have a sense of 'how much' studying they should be aiming for and what the benefits will likely be.

https://doi.org/10.1017/cnj.2024.24 Published online by Cambridge University Press

In sum, we think that linguists should be aware of the benefits of effortful retrieval as a study practice, and design courses that actively support students in engaging in this technique. While it is indeed more effortful, it likely has long-term benefits for improved student learning. Finally, we hope that the free, open-source software we have introduced here will assist instructors in creating randomized assessments that promote better learning.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/cnj.2024.24.

Acknowledgements. We gratefully acknowledge the contributions of various people who have made this work possible. First and foremost are our students, who have gamely gone along with all of our new ideas and provided us feedback along the way. We are also especially grateful to the audience at the 2021 meeting of the Canadian Linguistic Association / Association canadienne de linguistique, to the editors of this special issue on pedagogy, and to our colleagues who have provided feedback on this approach, especially Garrett Nicolai. Finally, we also recognize Gunnar Hansson and Douglas Pulleyblank in particular for providing us with information about their sections of LING 311.

References

Agarwal, Pooja K., Jeffrey D. Karpicke, Sean H. K. Kang, Henry L. Roediger III, and Kathleen B. McDermott. 2008. Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology* **22**: 861–876. https://doi.org/10.1002/acp.1391.

Brown, Peter C., Henry L. Roediger III, and Mark A. McDaniel. 2014. *Make it stick: The science of successful learning*. Cambridge, MA: The Belknap Press of Harvard University Press.

Carrier, Mark and Harold Pashler. 1992. The influence of retrieval on retention. *Memory and Cognition* **20**(6): 633–642. https://doi.org/10.3758/BF03202713.

Haudek, Kevin C., Luanna B. Prevost, Rosa A. Moscarella, John Merrill, and Mark Urban-Lurain. 2017. What are they thinking? Automated analysis of student writing about acid–base chemistry in introductory biology. *CBE-Life Sciences Education* **11**: 283–293. https://doi.org/10.1187/cbe.11-08-0084.

Hubbard, Joanna K., Macy A. Potts, and Brian A. Couch. 2017. How question types reveal student thinking: An experimental comparison of multiple-true-false and free-response formats. *CBE-Life Sciences Education* **16**(2): 1–13. https://doi.org/10.1187/cbe.16-12-0339.

Huxham, Mark, Fiona Campbell, and Jenny Westwood. 2012. Oral versus written assessments: A test of student performance and attitudes. *Assessment & Evaluation in Higher Education* **37**(1): 125–136. https://doi.org/10.1080/02602938.2010.515012.

Iannone, Paola, Christoph Czichowsky, and Johannes Ruf. 2020. The impact of high stakes oral performance assessment on students' approaches to learning: A case study. *Educational Studies in Mathematics* **103**: 313–337. https://doi.org/10.1007/s10649-020-09937-4.

Kang, Sean H. K., Kathleen B. McDermott, and Henry L. Roediger III. 2007. Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology* **19**(4–5): 528–558. https://doi.org/10.1080/09541440601056620.

Karpicke, Jeffrey D. and Henry L. Roediger, III. 2008. The critical importance of retrieval for learning. *Science* **319**(5865): 966–968. https://doi.org/10.1126/science.1152408.

Martinez, Michael E. 1999. Cognition and the question of test item format. *Educational Psychologist* **34**(4): 207–218. https://doi.org/10.1207/s15326985ep3404_2.

Meyer, Anne, David H. Rose, and David Gordon. 2014. *Universal design for learning: Theory & practice*. Wakefield, MA: CAST Professional Publishing.

Momsen, Jennifer, Erika Offerdahl, Mila Kryjevskaia, Lisa Montplaisir, Elizabeth Anderson, and Nate Grosz. 2013. Using assessments to investigate and compare the nature of learning in undergraduate science courses. *CBE-Life Sciences Education* **12**: 239–249. https://doi.org/10.1187/cbe.12-08-0130.

Nedjat-Haiem, Matthew and James E. Cooke. 2021. Student strategies when taking open-ended test questions. *Cogent Education* **8**(1): 1877905. https://doi.org/10.1080/2331186X.2021.1877905.

Rawson, Katherine A. and John Dunlosky. 2011. Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General* **140**(3): 283–302. https://doi.org/10.1037/a0023956.

Roediger, Henry L., III and Jeffrey D. Karpicke. 2006. *The power of testing memory: Basic research and implications for educational practice*. Perspectives on Psychological Science **1**(3): 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x.

Roediger, Henry L., III and Elizabeth J. Marsh. 2005. The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **31**(5): 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155.

Roelle, Julian and Kirsten Berthold. 2017. Effects of incorporating retrieval into learning tasks: The complexity of the tasks matters. *Learning and Instruction* **49**: 142–156. https://doi.org/10.1016/j.learninstruc.2017.01.008.

Rowland, Christopher A. 2014. The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin* **140**(6): 1432–1463. https://doi.org/10.1037/a0037559.

Scouller, Karen. 1998. The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education* **35**: 453–472. https://doi.org/10.1023/A:1003196224280.

Sehyr, Zed Sevcikova, Naomi Caselli, Ariel M. Cohen-Goldberg, and Karen Emmorey. 2021. The ASL-LEX 2.0 Project: A database of lexical and phonological properties for 2,723 signs in American Sign Language. *The Journal of Deaf Studies and Deaf Education* **26**(2): 263–277. https://doi.org/10.1093/deafed/enaa038.

Stark, Philip B. and Richard Freishtat. 2014. An evaluation of course evaluations. *ScienceOpen Research* **0**: 1–7. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1.

Struyven, Katrien, Filip Dochy, and Steven Janssens. 2005. Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education* **30**(4): 325–341. https://doi.org/10.1080/02602930500099102.

Supriya, K., Chris Mead, Ariel D. Anbar, Joshua L. Caulkins, James P. Collins, Katelyn M. Cooper, Paul C. LePore, Tiffany Lewis, Amy Pate, Rachel A. Scott, and Sara E. Brownell. 2021. Undergraduate biology students received higher grades during COVID-19 but perceived negative effects on learning. *Frontiers in Education* **6**. https://doi.org/10.3389/feduc.2021.759624.

Theobold, Allison S. 2021. Oral exams: A more meaningful assessment of students' understanding. *Journal of Statistics and Data Science Education* **29**(2): 156–159. https://doi.org/10.1080/26939169.2021.1914527.

Theophilides, Christos and Omiros Dionysiou. 1996. The major functions of the open-book examination at the university level: A factor analytic study. *Studies in Educational Evaluation* **22**(2): 157–170. https://doi.org/10.1016/0191-491X(96)00009-0.

Thomas, Patrick R. and John D. Bain. 1984. Contextual dependence of learning approaches: The effects of assessments. *Human Learning* **3**(4): 227–240.

To, Jessica, Ernesto Panadero, and David Carless. 2022. A systematic review of the educational uses and effects of exemplars. *Assessment & Evaluation in Higher Education* **47**(8): 1167–1182. https://doi.org/10.1080/02602938.2021.2011134.

Waterfield, Judith and Bob West. 2006. Inclusive assessment in higher education: A resource for change. URL https://www.plymouth.ac.uk/uploads/production/document/path/3/3026/Space_toolkit.pdf.

Cite this article: Vesik K, Hall KC (2024). Improved student learning through active retrieval practice and random-sampled exams. *Canadian Journal of Linguistics/Revue canadienne de linguistique* **69**, 285–306. https://doi.org/10.1017/cnj.2024.24