The Memrise Prize, an International Research Competition: A Pragmatic Trial to Compare Methods for Learning Foreign Language Vocabulary.

Rosalind Potts [a]*, Gesa van den Broek [b, 1], Anke Marit Albers [c], Jan Balaguer [d], Ruud Berkers [c], Mario de Jonge [e, 2], Asif Dhanani, Paul K. Gerke [f], Alysha Jivani, Boris Konrad [c], Carolina E. Kuepper-Tetzel [g, 3], Mark A. McDaniel [g], Toshiya Miyatsu [g, 4], Nils C. J. Müller [h], Walter Reilly [g], Christopher Summerfield [d], Hannah Tickle [d], Sharda Umanath [g, 5], Marlieke van Kesteren [i], Ed Cooke [j], Ben Whately [j], Robert A. Bjork [k], Jude Weinstein-Jones, and David R. Shanks [a].

[a] Division of Psychology and Language Sciences, University College London, London, UK.

[b] Behavioural Science Institute, Radboud University Nijmegen, The Netherlands

[c] Donders Institute for Brain, Cognition, and Behaviour, Radboud University Medical Center, Nijmegen, The Netherlands

[d] Department of Experimental Psychology, University of Oxford, Oxford, UK.

[e] Department of Psychology, Erasmus University, Rotterdam, The Netherlands.

[f] Radboud University Medical Center, Nijmegen, The Netherlands

[g] Department of Psychological and Brain Sciences, Washington University in St Louis, USA

[h] Radboud University, Nijmegen, The Netherlands

[i] Department of Psychology, Stanford University, Stanford, Santa Clara, California, USA

[j] Memrise, London, UK.

[k] Department of Psychology, University of California, Los Angeles, USA.

Present affiliations:

[1] Utrecht University, Department of Education, The Netherlands

[2] ICLON Graduate School of Teaching, Leiden University, Leiden, the Netherlands

[3] School of Psychology and Neuroscience, University of Glasgow, UK.

[4] Florida Institute for Human and Machine Cognition, USA

[5] Claremont McKenna College, California, USA

Author Note

Correspondence concerning this article should be addressed to Rosalind Potts, Division of Psychology and Language Sciences, University College London, Gower Street, London WC1E 6BT. Email: rosalind.potts@ucl.ac.uk.

Word count

Main text = 15,944. Introduction and discussion = 2877. Supplemental materials = 9,767

Informed consent

The study was conducted in compliance with appropriate Institutional Review Boards (IRB). Informed consent was obtained from the participants. Full details are provided in the Supplemental Information (Appendix C).

Abstract

How well do learning techniques work in the real world, and what happens when several techniques are combined? We conducted a competition in which international research teams developed methods to maximize the number of correct translations that learners could acquire in 1 h and successfully recall 1 week later. Teams initially tested their method for learning 80 Lithuanian-English words pairs against a standardized control method. Five shortlisted methods and the control condition were then compared on a common online platform, using Lakota-English pairs, with retention data collected from over 3,803 users of an online learning tool. The winning entry, which combined a visual mnemonic technique with retrieval practice and an adaptive algorithm for introducing new words, achieved an average of 27.23, 95% CI [26.08, 28.38] out of 80 word pairs recalled. This work highlights the contribution that competitions can play in addressing practical questions about human learning and memory.

Keywords: competition; learning; memory; mnemonics; vocabulary; Memrise.

General Audience Summary

Various techniques such as spacing, retrieval practice, and the use of mnemonic strategies have each been shown to be effective at optimising learning under laboratory conditions, typically when employed on their own. How well do these techniques work in the real world, and what happens when several techniques are combined? Learners acquiring the vocabulary of a foreign language have a limited time budget and hence effective learning methods must adopt an optimal trade-off between the value of a technique and the time taken to implement it. We conducted a two-stage competition in which international research teams developed methods to maximize the number of correct translations that learners could acquire in 1 h and successfully recall 1 week later. In Stage 1, the teams tested their method for learning 80 Lithuanian-English words pairs against a standardized control method. Five shortlisted methods and the control condition were then compared on a common online platform in Stage 2, using 80 Lakota-English pairs, with retention data collected from over 6,000 users of an online learning tool. After exclusions, the final sample was 3,803. Retention declined with age and was higher in females, and participants showed some metacognitive insight into the durability of their learning. The winning entry, which combined a visual mnemonic technique with retrieval practice and an adaptive algorithm for introducing new words, achieved an average of 27.23 out of 80 word pairs recalled. This work highlights the contribution that competitions can play in addressing practical questions about human learning and memory.

The Memrise Prize, an International Research Competition: A Pragmatic Trial to Identify Effective Methods for Learning Foreign Language vocabulary.

We live in an age in which, through technology such as the internet, information is more accessible than ever before and opportunities to acquire new knowledge abound outside the traditional classroom. How can we best take advantage of these opportunities and ensure that learning proceeds with maximum efficiency? Much laboratory research has been devoted to identifying optimal learning techniques (Dunlosky & Rawson, 2019). A review evaluating the usefulness and generalizability of ten commonly employed study strategies (Dunlosky et al., 2013) concluded that there was strong evidence in support of some techniques, such as retrieval practice (testing) and spaced practice (multiple tests separated in time) as effective learning tools across a variety of situations, and much weaker evidence for other techniques, such as keyword mnemonics. The authors argued that although keywords could be useful for associating foreign language words with their translations, the time taken to generate them outweighed their usefulness by comparison with other techniques such as retrieval practice.

This raises a fundamental question: What happens when several techniques are combined? What combinations are effective and which ones lead to inefficient trade-offs? There has been some recent interest in exploring such questions (e.g., see Latimier, Peyre, and Ramus, 2021, for a meta-analysis of the emerging literature on combining spacing and retrieval practice, and McDaniel, 2023, for a review of studies combining mnemonic techniques with retrieval practice). Real-world learners have limited time budgets, so the question we posed was: If someone had an hour in which to study some new foreign language vocabulary, what would be the best use of that hour to ensure maximum recall a week later?

To address this question, two of us (RP and DS) adopted a novel approach to research in this field: We devised an international research competition, inviting learning researchers from across

the globe to submit their best solutions to the challenge we posed. The competition rules were designed to minimize constraints on the possible solutions, while also encouraging methods that would be generalizable beyond the materials and conditions of the competition. The goal was to gather the best ideas that contestants could come up with, drawing on their experience and knowledge of research in this field, and then pit those solutions against each other in a between-subjects experiment with a large participant sample. Laboratory research often focuses on a theoretical question or on determining the efficacy of a theoretically-motivated variable. Borrowing clinical trial terminology, here our focus was on effectiveness rather than efficacy. Whereas efficacy research asks whether a manipulation produces an effect under ideal, controlled conditions, effectiveness (pragmatic) trials seek to estimate the impact of an (often complex) intervention under real-world conditions (Porzsolt et al., 2015).

The competition comprised two stages. The first was a laboratory stage, in which research teams developed and tested their proposed solution in their own laboratories against a standardised control task, then submitted their data to our panel of judges (R. Bjork, J. Weinstein, R. Potts, and D. Shanks). For the second and final stage we collaborated with Memrise (www.memrise.com), creators of an online foreign language learning tool. The five most promising solutions identified from Stage 1 were implemented by the Memrise team on a common online platform to compete against each other and against the control task in a large-scale experiment, with Memrise users, people with an intrinsic interest in language learning, recruited as participants. By giving researchers free rein to come up with creative solutions to the problem posed, we hoped that their solutions would make use of a range of strategies and combinations of strategies that would help us identify what the key elements of a successful learning regimen might look like under conditions of limited study time. By advertising for participants from among the large Memrise user community, we hoped to achieve a large sample of participants interested in language learning. Whereas for the first (laboratory) stage the sole dependent measure was the final test score, for the second (Memrise) stage we included other measures: judgments of learning (JOLs), effectiveness and enjoyment

ratings following study and following final test, as well as demographic measures such as age, gender, and native language status.

By devising a competition in this way, we hoped to shed light on how learning techniques can be efficiently combined to foster durable learning. The competition format complements not only laboratory studies but also other language learning research conducted in applied settings (e.g., Bryfonski & McKay, 2019) and using crowdsourced samples (e.g., Shortt et al., 2021).

Overview

This article about the competition is organised in two parts, reflecting the two stages of the competition, and around four overarching questions. First we describe the competition methodology and give an overview of the solutions that were submitted as entries to Stage 1 of the competition, focusing particularly on those that were most successful at this stage. The question (Research Question [RQ] 1) we were interested in here was: What strategies did the participating memory researchers choose to include when designing their optimal learning solutions? What elements did the successful solutions have in common and what elements were unique to particular solutions? Overlap between solutions may reflect adoption of widely accepted techniques, whereas differences between them point to strategies on which prior research provides less guidance.

Then we turn to the second stage (the Memrise run), describing the design of this stage of the competition and its outcomes. How did the five finalist solutions compare (with each other and with the control task) in terms of final test scores, subjective ratings that learners made about their learning experience, and learners' metacognitive judgments about learning success (RQ2)? Final test scores in this second stage were used to determine the competition winner. Third, we asked whether the effects of the different learning solutions in that second stage were moderated by participants' age, reported gender and native language status (RQ3). Fourth, we related selected study strategies employed by the finalist solutions and measures of learning during study to

outcomes on the final test, to elucidate the mechanisms that could explain differences between the solutions (RQ4).

Stage 1: Laboratory Stage

The aim of Stage 1 was to launch a competition inviting contestants to submit their best solutions to the challenge we posed and, from the entries received, to identify the most promising solutions for the Memrise team to implement as a between-subjects experiment in Stage 2. In this section we outline the competition rules, recruitment of contestants, materials supplied to contestants to be used for testing their solutions against our baseline condition, and details of the procedure for the baseline condition and the final test, both of which were designed and supplied by us. We conclude this section by summarising the entries we received from contestants, the judging process, and key features of the entries that were chosen to be represented in Stage 2, the Memrise run.

Method

Competition Rules and Design

The task parameters were designed by the UCL authors (RP and DS) and agreed with Memrise Chief Executive Officer Ed Cooke and Chief Operating Officer Ben Whately. Researchers entering the competition were invited to develop a solution and test it in their own laboratories, in a between-subjects experiment, by comparing it with a baseline study condition (the control task) that was developed by the UCL authors and supplied to contestants in the form of a compiled executable program. Contestants were allowed to present their experimental condition in any form they chose (e.g., computer program, PowerPoint, pen and paper etc). The competition rules stipulated that the experiment was to consist of two phases, an hour-long study phase and a test phase, and that the study hour was to take place within a single session, with the test phase - a self-paced cued recall

test, developed and supplied by us - taken seven days later. See SI for the full rules as well as further details on all aspects described in this section.

Recruitment and Contestants

The competition was advertised as being open to professional researchers and non-researchers alike and was publicized via the Memrise website, other websites (UCL, the Psychonomic Society, the American Psychological Association (APA)), in various press outlets, through social media, and by email to individual researchers and research groups working in the field of optimal learning techniques. A Facebook page was set up for researchers to share ideas and post queries, and for us (RP and DS) to post updates on the competition.

Participants

Contestants were responsible for recruiting their own participants in Stage 1. These ranged from MTurkers to friends and family. See SI Table 1 for details.

Materials

Stimuli, which were provided to contestants in an Excel file, consisted of a list of 80 Lithuanian-English word pairs (e.g., *arbata-tea*) selected from a larger set normed by Grimaldi et al. (2010). See Appendix A for the list. Only nouns were chosen, with the intention of aiding memorisation by using words that could be easily imagined. A control task for the study phase, and a cued recall test for the final test, both programmed in Visual Basic by the first author (RP) and supplied as compiled executable computer programs that could be run on any PC without the need for specialist software, were available for downloading from the Memrise website, as was the stimulus list. Contestants were told that they could present their experimental method in any way that they chose, while they should use the programs provided for the control task and final test. However, we allowed contestants who wanted to run their experiment online to create their own versions of the control and final test programs, as long as these were approved by us. A link to one

version, by Asif Dhanani and Alysha Jivani, was posted on the Memrise Prize Facebook page for others to use if they wished. Unfortunately, this came a little too late for some potential contestants. The increase in the availability of experiment building software in recent years would make it much easier to run an online version of such a competition now.

Procedure

Study Phase: Control task. All 80 Lithuanian-English word pairs were displayed one by one, in randomised order, on the computer screen, for participants to study for as long as they chose. On the left of the screen, a counter kept track of which study cycle (i.e., iteration of the complete set of words) and which item within that cycle had been reached, starting with Cycle 1, Word 1. Participants clicked a "Next" button when they were ready to move on to the next item. At this point, the word pair disappeared from the screen for 500ms before the next word pair was displayed. The counter always remained on screen and was updated when the new word pair appeared. Once all 80 word pairs had been studied in this way, they were presented again in a new random order. This process was repeated until an hour had passed.

Study Phase: Experimental Task. Each contestant had complete freedom to design and implement the experimental condition as they chose and run it against the control task in their own laboratories. SI Table 1 shows the key features of the entries, including how they were implemented and presented, the recruitment method and number of participants, and the main strategies appearing in each solution. Detailed descriptions of the procedures used in the five solutions that were chosen to go forward to Stage 2 are included in the Stage 2 Method section. Below we outline the key common and unique features of these five solutions. Further details of the procedures used in the remaining solutions can be obtained from the first author on request.

Final test phase. In the final test, administered seven days after study in both control and experimental conditions, all 80 Lithuanian cues were presented one by one in randomised order. The participant's task was to recall and enter the English translation. Participants had unlimited time in

which to make a response, and did not receive feedback on their responses, neither were they able to return to a previous item. When all 80 cues had been tested, participants were given the option to view their score. This score was based on strict scoring (the response exactly matched the target) but contestants were asked to provide both a strict score and a lenient score (response differed from the target by no more than two letters) when they submitted their entries.

Outcomes: Judging of the Competition Entries

In this section, we give an overview of characteristics of the entries submitted by contestants and of the judging process used to decide which solutions would be implemented in the Memrise run of the competition (Stage 2). Further details can be found in the SI.

Overview of the Stage 1 entries and shortlisting process

Thirteen entries were received from eleven groups, of which seven were from the USA, two from the Netherlands, one from Poland and one from the UK. Of these, six were research groups operating within a university environment. Two of the teams submitted data for more than one version of their solution. (In both cases these were compared against the same control group.) See SI for details of how entries were submitted and the information contestants supplied at this stage.

Many of the experimental tasks used common strategies, such as retrieval practice, keyword mnemonics, and learning algorithms. Eight of the groups used computerised tasks, with three solutions being presented via PowerPoint slides. Table SI1 summarises key features of the entries. For each entry, the judges considered whether the rules of the competition had been adhered to (one solution was excluded on this basis), as well as the size of the effect achieved by the experimental method over the control. See SI for details of the judging process and decisions, and SI Table 2 for the means and effect sizes. This process yielded a shortlist of five solutions. We had originally planned to select just three solutions to go forward to Stage 2 but, after analysing the shortlisted solutions, the Memrise team generously offered to program all five solutions for Stage 2.

We describe the finalists' solutions in more detail below, addressing our first research question, before turning to Stage 2, the Memrise run.

The Five Finalist Solutions: Common and Unique Features

What elements did the successful solutions have in common and what elements were unique to particular solutions (RQ1)? The five finalist solutions differed in several ways but also had features in common, some of which are summarised in Table 1. For convenience, we have given each solution a label reflecting its most salient features (see column 1 of Table 1). A brief description of each solution can be found in the second column of Table 1. Note that detailed descriptions are provided in the Stage 2 Method section, and illustrations are provided in the SI. We summarize important overlapping and distinctive features below.

Instructions. For most solutions, instructions were brief, involving short written explanations about the experiment and suggesting possible mnemonic techniques (e.g., keyword method) to use. Two solutions (*Link Phrases* and *Memory Champion*) began with an instruction video that explained such methods more extensively using visual aids. The video of the *Memory Champion* solution involved a memory champion (one of the researchers) explaining the method of loci and assuring participants that using it in the way instructed would substantially help them to learn the words.

Retrieval practice. All solutions employed retrieval practice, i.e., presenting a cue (the foreign word) to which the participant was to respond by recalling the correct target (the English translation), though this was employed to a different extent across the five solutions. These cued recall tasks always involved overt retrieval, with the participant typing the response. One solution, *Memory Champion*, also included a covert free recall task at two points in the study hour: Participants were shown background images they had seen during study and were asked to recall, without typing their responses, all the items they had studied against that background, with no cues present.

Adaptive Learning. Three solutions (*Study-Test*, *Errorful Generation* and *Memory Champion*) used adaptive learning algorithms to determine the timing and number of item presentations. These algorithms presented new items for study only when previously presented items had been learned to a certain criterion. This meant that, for these solutions, it was possible for participants to complete the study phase without having seen all 80 items in the stimulus set, depending on their learning rate. These solutions computed adaptive weights for each word, depending on how well they were remembered and the interval between presentations. These weights were used to determine how often a word was presented throughout the experiment, to allow harder-to-learn words to appear more often.

Keywords. Two of the solutions (*Mediators* and *Link Phrases*) instructed participants to generate a mediating keyword or phrase connecting the cue and the target and to enter this on the computer when prompted. For these solutions, the foreign word was always presented simultaneously with its translation on initial presentation and participants were to generate a keyword in response to the foreign-English pair. In *Memory Champion*, the foreign word was first presented without its corresponding translation and participants were encouraged to generate a keyword and associated image related to the foreign word *before* seeing the translation, then to link the image they had created with the meaning of the word when the translation was revealed. For example, a participant might respond to the Spanish word “zumo” by imagining a sumo wrestler (keyword). Then, on seeing the translation “juice”, they were to imagine the sumo wrestler drinking juice. A fourth solution (*Errorful Generation*) suggested the keyword method as an optional strategy, but neither this one nor *Memory Champion* required the keyword to be entered into the computer.

Trial Sorting. One solution, *Study-Test,* sorted the stimulus set at the start of the study phase according to several orthographic properties, so that all participants studied the items in the same order. (Further details are given below.) All five other solutions presented the items in an order randomised on a per participant basis.

Batching. One solution (*Link Phrases*) presented cue-target pairs for study in batches of 10 before testing them by either multiple choice test or matching test. For the matching test, the 10 cues and 10 targets were presented in separate columns and the participant's task was to match the cues with the correct targets. After every 40 words, all 40 words were given retrieval practice, i.e., tested via cued recall. This design meant that participants should typically encounter all 80 items. In another solution (*Mediators*), participants studied all 80 cue-target pairs before engaging in several blocks of retrieval practice, each comprising all 80 items. Finally, the *Memory Champion* solution batched words in groups of six and linked each group to a background image of a room, presenting the cue against the corresponding room image on first presentation and following each unsuccessful retrieval attempt. Ten room images were used in all. After 60 items had been learned, further items were introduced and matched with previously seen room images. These study and retrieval rounds were interleaved with two covert free recall tasks in the middle and at the end of the study phase.

Cue before translation. Two solutions (*Errorful Generation* and *Memory Champion*) presented the foreign word alone before presenting the translation during the study phase. In the case of *Errorful Generation*, participants were encouraged to enter a guess as to the meaning of the word before viewing the translation, while for *Memory Champion* they were to generate, but not enter, a potential keyword and associated image. These two solutions differ from *Study Test, Mediators* and *Link Phrases* in that the participant's response (guess or keyword) is generated before the target has been seen and therefore before the meaning of the word is known.

Visual presentation style. *Memory Champion* presented items against a backdrop of images of rooms during the initial encoding trials and in feedback after incorrect retrieval trials. Participants were instructed to visualise their self-generated keyword interacting with the word's translation in that room. Partway through the study hour, images of the rooms were presented again, without cues, and the participant was instructed to recall, covertly, the items studied in that room.

The Memrise team implemented the five finalists' solutions on their platform alongside the control condition, enabling us to compare them in a between-subjects experiment with six groups in Stage 2 of the competition, which is reported next.

Stage 2: Memrise stage

The aim of Stage 2 was to determine which solution yielded the highest final test score when they were directly compared with one another and with the control condition in a between-subjects experiment with random allocation of participants to solutions. The basic format of the experiment was the same as in the laboratory stage and involved a one-hour study phase during which participants studied up to 80 foreign (Lakota-English) vocabulary items, followed by a cued recall test one week later. We explored some additional dependent variables in this stage: As well as final test scores for each of the six conditions, we collected judgments of learning (JOLs) following the study phase; specifically, we asked participants to predict how many words out of 80 they would remember when tested one week later. JOLs provide a standard measure of metacognitive awareness about memory durability (Rhodes, 2016). We also took measures of effectiveness and enjoyment, each on a 5-point scale, both at the end of the study phase and after the final test, and we asked participants to report their age in years, their gender (female/male/other), and whether English was their native language. Our hypotheses were as follows.

Hypotheses

Final Test scores, JOLs, Effectiveness and Enjoyment measures

We hypothesised that the six groups (the five experimental solutions plus the control condition) would differ in final test score, JOLs, and effectiveness and enjoyment ratings. If the outcomes of Stage 1 were to be replicated, we would expect that the *Memory Champion* solution would emerge the competition winner, achieving the highest final test score, with *Errorful Generation* close behind. We requested JOLs (judgments of how much information will be

remembered) as well as effectiveness ratings (judgments about how good the technique is perceived to be) as these are potentially distinct. A wealth of previous literature has found dissociations between participants' JOLs and their actual test performance (e.g., Kornell et al., 2011; Rhodes, 2016), so we expected that we might see such dissociations in our data. For example, a solution that feels easy and fluent to participants might evoke high metacognitive ratings but actually be relatively ineffective (e.g., Kirk-Johnson et al., 2019). Enjoyment ratings and their alignment with actual effectiveness are also important, as learners might shun an effective but unenjoyable technique.

Effect of Age and of Native Language Status

Based on previous literature, we expected to see a decline in final test scores with age (e.g., Ward et al., 2020) and an advantage for native English speakers over non-natives in final test performance, since the targets were more familiar for the former (Hall, 1954).

Effect of Gender

Previous literature has suggested an advantage for females over males in verbal memory tasks, with a recent meta-analysis (Asperholm et al., 2019) finding an advantage of Hedges' g = 0.28. In line with the general findings of this literature, we predicted a small recall advantage for females over males. There is some evidence that women tend to give lower confidence ratings to their performance on certain cognitive tasks than men, even when actual performance is equal (e.g., González-Betancor et al., 2019; Pallier, 2003). We therefore predicted that females' JOLs would exhibit underconfidence relative to males'.

Method

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Design

The six solutions (five finalists' tasks and the control task) were compared with one another in a between-subjects design, with Solution as the independent variable with six levels. Participants were randomly assigned to conditions. The main dependent measures were final test score, JOL, effectiveness ratings immediately following study and immediately following final test, and enjoyment ratings at these same time points. Other measures for which we had hypotheses were age, gender and native language status. We also collected data on highest educational level, number of years in formal education and languages known other than English, for use by future researchers. We have not analysed these data for the current study.

Participants

We aimed to recruit as many participants as could be recruited during the approximately six-month period that the experiment was available. The experiment was initially advertised to participants via a blog on the Memrise website, followed by a pop-up notification on the Memrise system and mass emails to Memrise users, with a link to a sign-up page. Participants could win an iPad by participating in the competition. A total of 43,654 participants filled in their email address on this sign-up page. Signing up triggered an automatic invitation email with a link to the study phase, which could be used once. If the link had not been used one week later, a second email was sent. Participants who completed the study phase received a "pre-final test" email 5.5 days later, telling them to expect an email 24 hours later with a link to the final test. The "final test" email was sent approximately 6.5 days after completion of study, with a link to the test phase of the experiment, and could be used at any time after that, though participants were strongly encouraged to complete the test seven days after study. After exclusions (see below), a total of 3,803 participants remained in the data analysis.

Materials

Stimuli. To ensure that competitors' solutions would generalize beyond the Lithuanian-English stimulus set used in Stage 1, we chose a new language and new set of words, with a wider

range of parts of speech, which remained undisclosed until the start of Stage 2. Given that Memrise users were likely to come from a variety of language backgrounds and to have a variety of experience of different languages, it was important to try to control, as far as possible, for prior exposure to the materials to be learned. We addressed this by choosing a language that would be unlikely either to be known to participants or be related to languages known to participants: Lakota, a Siouan language spoken by approximately 6,000 people mainly living in North and South Dakota, USA. The Lakota stimulus set can be found in Appendix B and further details about how the stimuli were chosen can be found in the SI.

Questionnaires. On completion of the study phase, participants were presented with a post-study questionnaire which asked them to give a judgment of learning, i.e., to predict how many items they thought they would recall in the final test a week later, by entering a number from 0 to 80, and to rate their enjoyment of the method and how effective they perceived it to be, each on a scale of 1-5 (where 1 was lowest and 5 highest). They were also asked how many words they had written down while studying. A post-test questionnaire was presented at the end of the final test, which included the same enjoyment and effectiveness questions as asked following study, as well as asking whether participants had written down any words during study, or looked up any words between study and test, and whether they had previously taken part in the experiment, so that data from participants who had failed to observe the rules could be excluded from analyses. Although it was not possible for someone to start the experiment twice with the same email address, we could not prevent people from participating a second time with a different email address, so these questions served as an additional check. See SI Appendix D for the post-study and post-test questionnaires.

Procedure

Study Phase. On clicking the invitation link to the study phase, participants were randomly allocated to one of the five finalists' solutions or the control condition. All participants were

informed that this was a research study being conducted by University College London in collaboration with Memrise. They were told that it was open to anyone over 18, and they were assured of anonymity and confidentiality and asked to give their consent. They were told that they would have an hour in which to learn 80 words in Lakota and that they had been randomly assigned to one of the learning methods that were the subject of the research. It was explained that, as this was a memory test, they should not write any words down during the study phase. After entering their age, gender and if English was their native language, they could click a "Next" button to start the study phase of their allocated solution, at which point the timer started. If the age entered was under 18, participants were not allowed to proceed to the study. The procedure for the six groups was as follows. See the SI for further detail of each of the solutions and Appendix E in the SI for sample screenshots.

Control. The procedure for the control group's study phase was the same as in Stage 1.

Errorful Generation. The study phase began with written instructions that encouraged participants to try to create mental images or mnemonics to help remember the word pairs. As an example, participants were shown the Swahili-English word pair *wingu – cloud* and as a possible mental image it was suggested to imagine a cloud with gigantic birdwings. It was stressed that the more bizarre the mental images participants came up with, the better they would be able to remember the word pairs.

On each trial, including the first presentation of an item, the cue (the foreign word) was presented on its own and participants were instructed to enter the English translation. They could check whether they were correct by pressing "Enter", after which the correct translation gradually appeared on the screen. One by one the letters of the correct responses were shown on the screen with 100 ms in between each successive letter appearance. After the complete word had been uncovered, it remained on screen until the total feedback duration (4 s) had expired. Then, another

item from the list was sampled for presentation. Trials were self-paced, with an upper limit of 20 s for a response, to prevent participants from lingering too long on any single item.

An algorithm was used to determine which item was to be presented next. Items were initially sampled from the list in a random fashion throughout the experiment. Importantly, however, each item was assigned a weight and these item-weights changed depending on the participant's performance during the trials. All items started with a weight of 100 at the beginning of the learning phase, giving each individual word pair an equal chance (1/80) of being sampled. However, if a participant's response for an item was incorrect, the weight for that item was increased to 2000 (*factor1*). Thus, by increasing the weight of an item, the chance of that item being sampled for a subsequent representation increased dramatically.

Since participants were only ever presented with test trials, they were bound to get items wrong at the start. Therefore, increasing the weight of non-recalled items ensured rapid (short-lagged) re-presentation. As more unrecalled items entered an increased weight state, the likelihood of any new item being sampled for presentation dramatically declined. This resulted in focused retrieval practice of a subset of items at the beginning of the learning phase, but items never received a "massed" presentation. The shortest possible lag between any two consecutive presentations of the same item was one intervening item. As soon as an item was correctly recalled, its weight returned to the default value of 100.

Items that had already been recalled during a prior presentation were treated differently from previously unrecalled items. That is, items that had already been recalled once were given a weight of 50 after being recalled for the second or third time. In contrast, previously recalled items that were missed during subsequent second or third presentations were given a weight of 1000 (*factor2*). After an item had received three or more successful recalls, the item-weight was set to 1 in the case of a subsequent correct response. If an item was missed on any subsequent test trial, the item weight was set to 10.

Lastly, because well-learned items received increasingly smaller weights, the average weight of all the items in the list also declined as learning progressed. To compensate for the decline in average item-weight, the factor-values used to determine the weight of unsuccessfully non-recalled items also declined during the experiment. For every 10 newly recalled items, 250 points were subtracted from factor1 (which had an initial value of 2000), and 125 weight-points were subtracted from factor2 (initial value: 1000). Factor1 and factor2 could never become zero. After all 80 items had been recalled at least once, factor1 and factor2 were both set at 100.

This algorithm resulted in a dynamically scaffolded learning schedule where, on average, the lag between any two subsequent presentations of the same item would expand as the number of successful retrievals for that item increased. However, for items that were answered incorrectly, the intervening lag was temporarily compressed by increasing the item's weight. The study phase ended when the hour was up.

Link Phrases. At the start of the experiment, participants were shown a pre-recorded video, lasting about 2 minutes, on how to use the "Link Word" method to remember word pairs. It instructed them to choose the first word that came to mind when they saw a foreign word and to create a phrase using both this word and the English translation of the foreign word. For example, for the word pair *stogas* - *roof*, a possible phrase could be "a toga party on the roof".

After the video, participants were presented with a foreign word and its translation above an input box where they could type their link word phrase. Participants were given 25 seconds to create and enter each phrase. After entering their phrase, they spent the remainder of the 25 seconds visualizing their phrase and imagining how it would look, feel, smell, etc, to engage multiple senses. In addition to word association, this method was intended to allow participants to add context and emotion to the words.

After every 10 words, participants were tested on the 10 words they had just learned, alternating between multiple choice and matching tests. During the multiple-choice rounds,

participants were given 6 seconds to select the correct English translation for the cue word from three possible options. After each response or when the time ran out, the correct answer was highlighted in green for 1.5 seconds and then the next multiple-choice question was presented. In the matching rounds, participants had 1 minute to match the current round's 10 cues with their English targets, following which the correct pairs were displayed.

After every 40 words, there was a timed cued recall test of the previous 40 words. Participants were presented with a cue word and were instructed to type its translation within 7 seconds. They could press a "show hint" button to display the link word phrase that they had created for the word pair during initial study. After entry of a response or when the time expired, the correct answer was shown. The hints allowed participants to practice their Link Word phrases while being re-exposed to the word pair.

Once all 80 words had been studied and tested in this way, there was a further cued recall test of items that had been incorrectly answered in either cued recall round. This test was self-paced and included the "show hint" button.

If there was time remaining after this, a three-column list of all 80 word pairs was presented on the screen. Each row displayed a cue word, its English translation, and the link word phrase the participant created for the word pair, and participants were instructed to study them in any manner that they chose until the study hour was up.

Mediators. In written instructions at the start of the study phase, participants were told that they would study 80 word pairs and then receive three opportunities to practice remembering the English translation of each word. Then the keyword method was explained: For each word pair, participants were encouraged to generate a keyword that was embedded within, or related to, the cue word and to form an image linking the keyword to the English translation. They were told that, when they came to be tested on the words, they should identify the keyword they had chosen, then recreate the image connecting the keyword and the English translation to help them recall the

English translation. Some examples were given to help participants to understand the method. Please see the SI for the exact instructions.

Following the instructions, participants studied all the 80 Lakota-English word pairs one by one in a random order. They were given 14 seconds for each trial but were able to advance to the next trial after 7 seconds if they thought they were ready by clicking on the "Next" button. After this initial encoding, participants took a one-minute rest during which they watched a video of a waterfall in a forest. They were instructed to imagine being there and to stand up and stretch. Next, participants engaged in a round of retrieval practice. They were shown the 80 Lakota words one by one in a new random order and were asked to retrieve and type in both the keyword they generated and the English translation. They were given 9 seconds for each trial but were able to advance to the next trial after 3 seconds if they finished typing. Feedback (showing both the Lakota and English words) was provided for 5 seconds after each trial. Participants took a one-minute rest break in the same way as after initial encoding.

Then, participants engaged in a second round of retrieval practice. They were given the 80 Lakota words one by one in a new random order and were asked to retrieve both the keyword they had generated and the English translation, but this time asked to type in the English translation only. They were given 4 seconds for each trial but were able to advance to the next trial after 2 seconds if they finished typing. Feedback (showing both the Lakota and English words) was provided for 2.5 seconds after each trial. This was followed by another one-minute rest. This second retrieval practice and rest break were then repeated until the study hour was up.

Memory Champion. Participants first watched an instruction movie in which a memory champion in a white lab coat demonstrated how to use a keyword mnemonic technique together with the method of loci. The memory champion explained that the technique had been developed in collaboration with neuroscientists and memory researchers. Participants were told that they would be shown an image of a room with a foreign word superimposed on it and were instructed to think

of a keyword, i.e., a word that they associated with the presented word due to phonological or orthographical similarity. They were instructed to picture this keyword within the scene, as vividly as possible. Then they were to press “Enter” to see the translation and were instructed to picture this translation in the scene, interacting with their keyword, again as vividly as possible. An example was given using the Spanish word “zumo” against a background showing a living room. A possible keyword could be “sumo”, so the participant was told to picture a sumo wrestler in the living room. When the translation, “juice”, appeared, they were to imagine the sumo wrestler drinking juice in the living room. Participants were encouraged to use this technique and to try using the background scenes as a visual aid, but they were also instructed that they should feel free to ignore the background scenes if they did not find them helpful.

Participants studied the foreign words in batches of six with a photo of a scene displayed simultaneously. Study trials were interspersed with retrieval trials. On each trial, the foreign cue was presented without its translation. On the first presentation of an item, there was no option to enter a response: Participants were to think of a keyword and then press “Enter” to see the translation. These study trials were always accompanied by an image of a room. The room image was at the top of the screen, with the cue word presented below it on the left and the participant’s cumulative score on the right. Subsequent presentations of that item were retrieval practice trials, in which a cue was presented without a background image and participants could enter the translation below the cue. The participant’s score was displayed on the right. If the response was correct, it turned green and the score increased by 10 points. If the response was incorrect, the incorrect response turned red and appeared crossed out and the correct answer was shown alongside in green for one second, together with the background image that was associated with the word. Trials were self-paced up to a maximum of 15s per trial.

An algorithm controlled the addition of new items and the number and spacing of repetitions of old items by calculating estimates of memory strength for each word as a function of

its practice history, i.e., the number and timing of earlier presentations and response accuracy at these presentations, and presented words before their memory strength fell below a specified practice threshold. This led to an expanding schedule of retrieval practice, as well as comparably more retrieval practice of "difficult" words than of easier words, and meant that participants who learned faster practiced more items.

For the background images, ten different photos were used (e.g., a living room, kitchen, gym, garden). The first six words were shown with the first image; the next six words were shown with the second image, and so on. When a participant managed to go through more than 60 words (10 rooms x 6 words per room), another two new words were added to each room. The rooms were always shown during the first presentation of a new word and were shown again when a participant failed to recall the correct translation of a word. Finally, if participants typed in the translation of a different word from the stimulus set, "smart feedback" was shown: A prompt, "You mixed up two words", was shown and after one second the cue word with the correct translation and the mixed-up word with its respective translation were displayed together.

Every 25 minutes, i.e., once about half-way through the training and once towards the end of the session, participants performed a free recall task. This was to provide some variety in the training to keep the participants motivated and alert, and to increase the chance that they could recall the words later. Participants were first asked to think of all the background images ("rooms") that they had seen during training. After 30 seconds or when the participants pressed Enter, they saw the rooms one at a time and were asked to recall all the items and associated visual imagery that they had created for the presented room. After 30s, or when the participant pressed Enter, all the words and translations for the room were displayed. They remained on screen until the participant pressed Enter, up to a maximum of 30s. If the study hour ran out during the free recall task, it was terminated. If a participant went through the free recall task quickly, retrieval practice continued after the second free recall phase until the study hour was up.

Study-Test. Participants were informed, in written instructions, that they would see Lakota-English word pairs, and that shortly after the initial presentation of a word pair, they would be asked to recall the translation from the Lakota cue. They were told that, in the case of an incorrect response, they would be shown the correct answer again and that, to begin with, they would see a few pairs of words repeatedly but that as their learning progressed, they would see more new pairs. Finally, they were encouraged to learn the pairs as well as possible, as they would be tested on them in a week's time.

Each item in the stimulus set was presented first as a single study trial, in which cue and target were presented together for participants to learn, and then as repeated, spaced retrieval trials, in which the cue alone was presented, and the participant's task was to enter the target translation. Cues were presented in red font and participants' responses appeared in black font. Trials were self-paced and were terminated when the participant pressed Enter, followed by an interstimulus interval of 1s before presentation of the next trial. In the case of a correct response on a retrieval trial, the next item was presented immediately. If participants entered an incorrect translation, they were shown the correct cue-target pair for 2s as corrective feedback.

The order and spacing of presentation of items was determined by an algorithm, which determined the trace strength and consolidation level of an item based on a combination of the participant's past performance (whether the participant had previously recalled that item) and the number of trials that had elapsed since the previous presentation, so this training schedule naturally spaced the word pairs and adjusted for the extent to which they had been learnt thus far. To begin with, a subset of the items was released for study and new items were introduced only as the older items became well learned. This meant that slower learners might not encounter all 80 items during the study hour. See the SI for a detailed description of the algorithm.

In addition, a procedure was applied at the start of study to rank the items from easiest to most difficult, using a model that had been developed to apply to any language (since the language

used for Stage 2 was not revealed to contestants until after the completion of Stage 1). This sorting was based on a wide range of characteristics of the items, such as the length of the words and the degree of similarity between the foreign and the English word. Presentation of easier items early in the study phase was done to ensure that even poor learners could master some of the items. Further details of the ranking model can be found in the SI. Presentation of items for study or retrieval practice continued until the study hour was up.

Post-study questionnaire. Following the study hour, participants in all conditions were presented with the post-study questionnaire. They entered their JOL and ratings of enjoyment and effectiveness and indicated how many words they had written down.

Final Test. Participants who completed the study phase could access the final test from their "final test" email, sent 6.5 days after they had participated in the study phase. The final test was identical for all participants and was the same as in Stage 1 of the competition, except for the use of the Lakota stimulus set instead of the Lithuanian-English word pairs. At the end of the test phase, participants were shown their score and they completed the post-test questionnaire. This included questions about their enjoyment of the method to which they had been allocated and its perceived effectiveness, each on a scale of 1-5 (where 1 was lowest and 5 highest). It also included questions designed to ascertain whether participants had written down any words during study, or looked up any words between study and test, and whether they had previously taken part in the experiment, so that data from participants who had failed to observe the rules could be excluded from analyses. Although it was not possible for someone to start the experiment twice with the same email address, we could not prevent people from participating a second time with a different email address, so these questions served as an additional check.

Data analyses

For RQ2, the comparison of solutions on learning outcomes and participants' ratings, the data were aggregated at participant level and compared in analyses of variance (ANOVAs), using *t*-tests for

contrast analyses. All *t*-tests use the Welch method. In addition, two-sided Bayesian *t*-tests (with a default Cauchy prior width of $r = .707$) were used to quantify the evidence for or against the null hypothesis, using the BayesFactor package (Morey & Rouder, 2022). Unless stated otherwise, we report the Bayes factor for the alternative hypothesis (BF_{10}). For RQ3, we conducted three two-factor analyses of variance to test whether participants' gender, age, or native language status moderated the effect of the experimental solution. For RQ4, the solutions were grouped based on various features (e.g., presence of adaptive spacing algorithm) to test whether these moderated differences in final test recall between the solutions. A generalized mixed effects logistic regression assessed the relationship between numbers of retrieval practice trials correct and incorrect at study and final scores. Only correctly spelt responses that were exact matches to the target were counted as correct for the purpose of calculating scores.

Results

We begin this section by reporting data on participation in the experiment at all stages from signing up to completing the experiment. Table 2 gives a detailed breakdown of progress for all invited participants for each solution. In summary, of over 43,000 people who signed up for the experiment, 13,473 (31%) started the study phase, of whom 6,028 (45%) completed it. Out of all participants who started the study phase, 5,243 (39%) went on to complete both phases of the experiment. Of those, 1,256 (24%) admitted to having written words down or looked words up during the course of the experiment, or to having done the experiment before. Their data were excluded from the dataset. Retention intervals between study and test phase for the remaining participants ranged from 6 to 32 days. To maximise the number of participants included in the dataset, while also maintaining an interval of approximately one week between study and test as stipulated in Stage 1, we allowed a retention interval of up to 9 days, yielding a total of 3,804 participants. One participant was removed from the dataset as inspection of their study data showed they had not experienced any trials after the instructions, leaving 3,803 participants.

Final Test Scores and Metacognitive Measures (RQ2)

We now address our second overarching question (RQ2): How did the five finalist solutions compare on final test scores and metacognitive measures? The competition rules were that the solution with the highest average final test score would be the winner. We were additionally interested in whether participants' subjective perceptions of the effectiveness of a solution were aligned with actual outcomes.

Final Test Scores. Figure 1 shows the final test scores, which differed significantly across solutions, $F(5, 3797) = 73.45$, $p < .001$, $\eta^2 = 0.09$. The *Memory Champion* solution achieved the highest score ($M = 27.23$, 95% CI [26.08, 28.38]), followed by *Study-Test* ($M = 24.79$ [23.53, 26.04]) and *Errorful Generation* ($M = 24.39$ [23.02, 25.76]). *Memory Champion's* advantage over all other solutions, including the runner-up, *Study-Test*, $t(1607.4) = 2.81$, $p < .005$, $d = 0.14$, $BF_{10} = 2.82$, was statistically significant. Notably, it achieved a mean score nearly double that of the *Control* group, $t(1245) = 15.05$, $p < .001$, $BF_{10} = 2.61 \times 10^{41}$, with a large effect size, Cohen's $d = 0.81$, confirming its effectiveness. Memory Champion was therefore declared the competition winner. A video demonstration of the Memory Champion method can be accessed at https://github.com/MrApplejuice/memprize-nijmegen/tree/1.0.

Judgments of Learning, Enjoyment and Effectiveness Ratings. Table 3 shows mean judgements of learning (JOL) and effectiveness and enjoyment scores taken at study and at test. In alignment with the retention data, *Memory Champion* also achieved the highest scores on all these measures. JOLs differed significantly across solutions, $F(5, 3797) = 70.24$, $p < .001$, $\eta^2 = 0.08$.

Relationship Between Participants' JOLs and Final Test Scores. Figure 2 shows the relationship between final test scores and participants' predictions (as percentages) for each solution. A mixed 2 x 6 ANOVA, with score as the dependent variable, Score Type (actual/judged [JOL] final test score) as a within-subjects factor, and Solution as a between-subjects factor, showed that JOLs ($M = 30.04$, $SD = 21.08$) were significantly higher than actual test scores ($M = 21.99$, $SD =$

17.40), F (1, 3797) = 601.93, $p < .001$, $\eta_p^2 = 0.14$, and that there was a significant interaction with Solution, F (5, 3797) = 64.61, $p < .001$, $\eta_p^2 = 0.08$. Only in the Mediators solution were test scores higher than JOLs. It is often found that participants' metacognitive judgments are poorly aligned with their actual test performance (Rhodes, 2016). In our study, at the group level, although JOLs were generally over-confident, the highest JOLs were given to the solution that yielded the highest score. While these results shed light on the group-level relationship between JOLs and final test scores, we also explored calibration at the participant level, within each solution, by calculating the absolute difference between JOLs and final test scores (i.e., regardless of whether the participant's JOL was over- or under-confident). The resulting calibration scores are *Control*: M = 17.01, SD = 15.12; *Errorful Generation*: M = 14.88, SD = 12.79; *Link Phrases*: M = 17.57, SD = 15.06; *Mediators*: M = 11.86, SD = 11.15; *Memory Champion*: M = 18.99, SD = 14.85; and *Study-Test*: M = 16.50, SD = 13.98. These differ significantly, F(5, 3797) = 18.49, $p < .001$, and reveal that calibration is best in *Mediators* and worst in *Memory Champion*. When data from all participants were considered, regardless of solution, there was a moderate but significant correlation between JOLs and final test scores, $r = .47$, $p < .001$, n = 3,803.

Effect of Age, Gender, and Native Language Status (RQ3)

Were the effects of the learning solutions moderated by participants' age, reported gender, and native language status?

Age. Reported ages ranged from 18 to 82. Considering all participants, and consistent with our hypothesis, final test scores tended to decline with age, as shown in Figure 3. An analysis of variance with solution and age as the factors found a strong effect of age, F(1, 3791) = 308.11, $p < .001$, $\eta_p^2 = 0.08$, while the main effect of solution continued to be significant as well, F(5, 3791) = 73.01, $p < .001$, $\eta_p^2 = 0.09$. The interaction was also significant, F(5, 3791) = 5.63, $p < .001$, $\eta_p^2 = 0.007$. *Memory Champion's* advantage over *Study-Test* and *Errorful Generation* narrowed somewhat as age increased, though this could be partly attributable to a floor effect.

Gender. More strikingly, there was a significant effect of gender on final test scores. The final dataset consisted of 2,001 participants self-reporting as female, 1,766 as male and 36 as other. As there were too few participants in the third category to draw meaningful conclusions, we report inferential analyses only for those self-reporting as female or male. Descriptive statistics for the 36 participants self-reporting as "other" yield a mean recall score of 28 (*SD*=19.3) and a mean JOL of 31.1 (*SD* = 20.5). For females and males, a 2 (Gender) x 6 (Solution) ANOVA, with score as the dependent variable, found that females (M = 23.99, SD = 17.75) achieved scores more than 4 points higher than males (M = 19.61, SD = 16.64), $F(1, 3755) = 70.48$, $p < .001$, $\eta_p^2 = 0.02$. Put differently, females recalled 22% more than males at final test ($BF_{10} = 3.65 \times 10^{11}$). The ANOVA also revealed a significant effect of solution, $F(5, 3755) = 74.53$, $p < .001$, $\eta_p^2 = 0.09$. Females outperformed males in every solution (see Figure 4; all $BF_{10} > 10$, except for *Control*, $BF_{10} = 1.46$, and *Memory Champion*, $BF_{10} = 0.85$). This gender difference was not reflected in participants' JOLs, however, where there was no significant difference between females (M = 29.58, SD = 20.76) and males (M = 30.55, SD = 21.44), $F(1, 3755) = 1.93$, $p = .164$, $\eta_p^2 = 0.00$, $BF_{10} = 0.10$. Lastly, the Solution x Gender interaction was not significant, $F(5, 3755) = 1.96$, $p = .081$, $\eta_p^2 = 0.00$.

Native language status. Participants were asked to report whether or not English was their first language. There were more non-native (N = 2,305) than native (N = 1,498) speakers in the final dataset. As we expected, native English speakers (M = 22.73, SD = 17.75) scored higher than non-native speakers (M = 21.51, SD = 17.16), $t(3119.5) = 2.09$, $p = .037$, Cohen's $d = 0.07$, but this effect was not supported by Bayesian analysis, $BF_{10} = 0.339$. There was no difference in JOLs between native (M = 29.50, SD = 21.19) and non-native speakers (M = 30.39, SD = 21.00), $t(3176.9) = 1.27$, $p = .203$, $BF_{10} = 0.084$, Cohen's $d = 0.04$. There was no interaction between native language status and Solution for either recall scores, $F(5, 3791) = 1.23$, $p = .291$, $\eta_p^2 = 0.00$, or JOLs, $F(5, 3791) = 1.10$, $p = .359$, $\eta_p^2 = 0.00$.

Relationship Between Study Strategies and Measures of Learning (RQ4)

We now explore the relationship between selected study strategies employed by the finalist solutions and measures of learning during study and final test outcomes (RQ4). Clearly, the research competition approach we took meant that there were many factors that differed between the solutions, so some of our analyses must necessarily be exploratory. In interpreting the results reported below, readers should bear in mind that the factors were not independent from each other, so confounds are likely to exist and be unaccounted for in statistical analyses comparing study strategies and learning outcomes.

We looked at the data in two main ways. First, we explored common features between the study methods and the effect of these on test scores and metacognitive measures. Next, we explored relationships between measures of learning during the study phase and final test outcomes. We begin by assessing the impact of chosen study strategies on test scores and metacognitive measures.

Number of Items Studied. Although all the solutions used the full stimulus set of 80 items, the use of adaptive learning algorithms in some solutions and self-pacing of study in others meant that not all participants encountered all 80 items during the hour-long study session. The number of distinct items encountered ranged from 12 to 80 across the whole dataset. For four solutions (*Control*, *Errorful Generation*, *Link Phrases*, *Mediators*) all or nearly all 80 items were studied on average. S*tudy Test* participants studied an average of 75.1 items. *Memory Champion* participants encountered the fewest unique items (M = 72.0) while achieving the highest final test recall, suggesting there may be some advantage to studying fewer items. The encoding of study items in this solution plainly yields paired-associate memories that are sufficiently enduring to compensate for their being relatively fewer in number compared to the other solutions.

Use of retrieval practice. Figure 5 shows a breakdown of trial types used in the five finalist solutions.[1] It is strikingly evident that the most common type of trial used in the three most successful solutions was retrieval practice with feedback on incorrect responses, a technique that, by comparison, was barely used in the less successful solutions. This extensive use of retrieval practice enabled the implementation of adaptive learning algorithms based on the accuracy of retrieval during practice trials in these three top-performing solutions. We now consider the effect of this strategy, and of two others that were also used in the more successful solutions: use of a keyword, and presentation of the cue alone before presentation of the target. See Table 1 for details of which solutions used each of these strategies and how they were implemented.

Adaptive Learning Algorithm. The use of adaptive learning algorithms in *Memory Champion*, *Errorful Generation* and *Study-Test* meant that the number of items encountered during study was determined by the learner's pace of acquisition. These three solutions all yielded significantly higher test scores than any of the other three solutions. It is no surprise then that, amalgamating data across solutions, a comparison of data from participants who studied with a method that included an adaptive learning algorithm (M = 25.57, SD = 17.82) with those who studied without such an algorithm (M = 16.30, SD = 15.05) revealed a significant and medium-to-large effect of the algorithm approach, $t(3491.6) = 17.22$, $p < .001$, $BF_{10} = 8.99 \times 10^{55}$, Cohen's $d = 0.56$.

Of course, there are several possible features of the adaptive learning algorithms that could have been responsible for this benefit to learning. Inclusion of an algorithm typically meant that, in cases where learning was proceeding more slowly, fewer than 80 words were encountered at study. It also meant that items were typically tested not long after being presented, whereas in the non-algorithm solutions they were first tested after either 40 items (*Link Phrases*) or all 80 items had been presented (*Mediators* and *Control*). The gradual introduction of new items in the algorithm

[1] In the SI we describe a software error that affected the recording of a small amount (0.17% across all solutions) of study phase data, particularly for Link Phrases participants, so these should be interpreted with caution.

solutions meant that learning proceeded at the pace of the learner rather than being experimenter-controlled, and that item presentations were spaced according to an expanding test schedule (Yan et al., 2020). Finally, as has been highlighted, these three solutions also made much more extensive use of retrieval practice than the other solutions. As we do not have a solution in which extensive retrieval practice is used with no adaptive learning algorithm, we cannot determine whether the benefit of these three solutions derives from the use of retrieval practice, the use of an adaptive learning algorithm, or a combination of the two. We return to this issue later.

Use of Keyword or Mediator Strategy. Several solutions either required or encouraged participants to use a keyword or mediator strategy. Two solutions explicitly required entry of a link phrase, keyword or mediator on presentation of the cue-target pair (*Link Phrases* and *Mediators*), while two encouraged participants to adopt such a strategy as a means of associating cue and target if the participant so chose (*Errorful Generation* and *Memory Champion*). The remaining two solutions, *Study-Test* and *Control*, made no mention of any kind of keyword strategy.

We grouped the study methods according to whether a keyword strategy was required, simply encouraged, or not mentioned at all. An ANOVA revealed a significant difference in final test score between these three categories, $F(2, 3800) = 79.19$, $p < .001$, $\eta_p^2 = 0.04$. When keywords were encouraged but not required ($M = 25.97$, $SD = 17.78$), scores were significantly higher than when no mention was made of them at all ($M = 20.43$, $SD = 17.58$), $t(2798.1) = 8.36$, $p < .001$, $BF_{10} = 3.27 \times 10^{13}$, Cohen's $d = 0.31$, and higher than when keyword entry was required ($M = 17.57$, $SD = 14.93$), $t(2228.7) = 12.63$, $p < .001$, $BF_{10} = 2.48 \times 10^{29}$, Cohen's $d = 0.51$. Solutions that made no mention of a keyword strategy also yielded higher test scores than those that required one, $t(2175.9) = 4.16$, $p < .001$, $BF_{10} = 161$, Cohen's $d = 0.18$. Although it might be tempting to conclude that the time taken to generate and enter keywords was at the expense of more fruitful strategies, such as retrieval practice, it is also the case that the solutions that required keyword entry (*Link Phrases* and *Mediators*) were solutions where all 80 items were always encountered and large batches (40 or 80)

of items were presented before any retrieval practice took place. The solutions that encouraged, but did not require, keywords or mediators were *Errorful Generation* and *Memory Champion*, but of course it is not possible to know to what extent participants adopted the strategy, which was strongly encouraged in the *Memory Champion* solution (in an introductory video in which a memory champion demonstrated the technique) and merely suggested in the *Errorful Generation* solution. However, these two solutions share another characteristic that sets them apart from the other solutions, to which we turn now.

Cue Presented Alone Before Presentation of Target. In two solutions, *Errorful Generation* and *Memory Champion*, the Lakota cue was presented on its own, without the target. In the case of *Errorful Generation*, the participant was encouraged to guess the translation while, in *Memory Champion*, to imagine something that the cue reminded them of (Pan & Sana, 2021; Potts & Shanks, 2014). Following this, the target translation was revealed. We compared final scores for these solutions ("cue alone") with scores for those where cues were always presented together with their targets on first appearance ("cue-target"). The "cue alone" solutions (M = 25.97, SD = 17.78), significantly outperformed the "cue-target" solutions (M = 19.24, SD = 16.59), $t(3188.4) = 11.78$, $p < .001$, $BF_{10} = 6.66 \times 10^{28}$, Cohen's $d = 0.39$. Interestingly, and at variance with some laboratory research (Potts & Shanks, 2014), the "cue alone" solutions also yielded higher JOLs, (M = 34.44, SD = 21.98 vs M = 27.00, SD = 19.88), $t(3119.6) = 10.67$, $p < .001$, $BF_{10} = 5.36 \times 10^{23}$, Cohen's $d = 0.36$, suggesting that active generation encouraged participants to believe they would do better at final test than when they had simply undergone retrieval practice.

Now we turn to relationships between measures of learning during the study phase and final test outcomes, exploring the relationship between retrieval practice opportunities and their outcomes at study, and final test scores.

Number of Retrieval Practice Trials Experienced at Study. The three best-performing solutions, *Memory Champion*, *Errorful Generation* and *Study-Test,* involved multiple spaced retrieval

practice trials per item. How did the number of retrieval practice (RP) trials affect final test outcome and did it matter to what extent retrieval practice was successful at study? The number of RP trials per studied item varied both within and between participants, according to the solution to which the participant was allocated and the point at which the item was introduced during the study phase. Some solutions (*Mediators*, *Link Phrases*) involved relatively little retrieval practice and the number of RP trials was similar for all items. For the three solutions with adaptive learning algorithms, the number of RP trials for an item was partly determined by how early in the sequence the item was introduced (items introduced earlier had more opportunity to receive retrieval practice) and partly by how quickly the participant learned the item (items that elicited more incorrect responses were subject to more retrieval practice). Furthermore, in each of these solutions, an incorrect response was always followed by a cue-target presentation, so the more incorrect responses there were for an item, the more encoding trials there were for that item.

In the following three sections, we begin by exploring how often a retrieval practice trial resulted in a correct or incorrect answer at study for each of the five experimental solutions, independently of final test score. We then relate those data to final test scores to determine if there is a relationship between the number of correct and incorrect retrieval practice trials at study and final test score. Finally, we look at the overall number of retrieval practice trials at study, regardless of whether participants responded correctly or incorrectly at study, and explore whether there is an optimum number of retrieval practice trials at study by relating the number of retrieval practice trials each item received to the final test score for that item and determining what percentage of items practiced that number of times were subsequently correct at final test. This information could potentially be used to recommend optimal study strategies.

How Often was a Retrieval Practice Trial Answered Correctly or Incorrectly at Study? This question is interesting because it gives us a measure not only of how much use was made of retrieval practice in a given solution but also of how quickly the solution enabled participants to

acquire new vocabulary during study. Figure 6 shows the mean number of RP trials that each participant underwent during the study phase in the five experimental conditions, split according to whether they were correct or incorrect at study. An item was counted as "correct" if the participant's response exactly matched the expected answer. (There were no RP trials, of course, in the *Control* condition.) For *Errorful Generation*, where the first presentation of an item required the participant to respond to the cue alone, this first trial for each item was a guess rather than a genuine retrieval practice opportunity, i.e., an opportunity to practice something that has already been learned. For this reason, the first presentation of each item in the *Errorful Generation* solution is not included in the data in Figure 6. There are several points of note. First, the three adaptive algorithm solutions made much more use of retrieval practice than the two lower performing solutions. However, of those three solutions, *Memory Champion*, which produced the highest overall final test score, had a lower mean number of RP trials than its two nearest rivals, suggesting that the sheer number of retrieval practice trials was not the only factor contributing to the success of this method. Particularly interesting is the comparison between *Study-Test* and *Errorful Generation*. These two solutions are very similar, in that they consist of repeated spaced retrieval practice on an expanding schedule, with *Study-Test* beginning with a cue-target presentation and *Errorful Generation* beginning with a cue alone and prompt to guess the target. They produced almost identical final test scores but the number of correct RP trials at study was substantially higher in *Study-Test*, suggesting that correct responding at study is also not the only important factor.

Relationship Between Accuracy at Study and Final Test Scores. What was the relationship between numbers of RP trials correct and incorrect during the study phase and final test scores one week later? A generalized mixed effects logistic regression analysis combining data across all solutions, with number of correct and incorrect RP trials at study as fixed effects and random intercepts by participants and items, revealed that final test score increased with the number of correct RP trials at study (b = 0.46, SE = 0.005, p < .001) and decreased with the number of incorrect RP trials at study (b = -0.15, SE = 0.004, p < .001). Moreover, this pattern held across all solutions,

with the exception that for *Link Phrases* (*b* = 0.06, *SE* = 0.046, *p* = .22), the effect of number of incorrect RP trials at study was not significant. Since incorrect RP trials at study provide re-exposure to the items via feedback, yet have a negative association with final recall, it follows that the beneficial effects of correct RP trials are attributable to retrieval rather than simply re-exposure, in line with laboratory research (Yang et al., 2021).

Relationship Between Number of Retrieval Practice Trials and Final Test Outcome. Here we explore the relationship between the number of retrieval practice trials for a given item, regardless of accuracy, and final test outcome. This issue is important from a design perspective because the researcher has no control over whether responses on RP trials will be correct or incorrect and can only program the overall number of such trials. We hence address the question "how many RP trials should be included for optimal final recall, under the constraints imposed?" For each participant, we determined the number of retrieval practice trials that each item underwent at study (the item's *RP level*) and whether or not that item was successfully retrieved at final test. We then calculated, for each participant, the percentage of items practiced at each RP level that were subsequently correctly retrieved at final test. These data are shown in Figure 7. It is interesting to note that, for *Study-Test,* while increasing numbers of retrieval practice trials initially boosted eventual recall, there were items that were practiced more than twenty times and still not retrieved at final test, despite the provision of feedback. This was not the case for the other solutions. Plainly, the time *Study-Test* participants spent trying to master these intractable items could have been put to better use.

General Discussion

We ran an international research competition to address the practical question that motivated this work – if a person has an hour in which to study some new foreign language vocabulary, what would be the best use of that hour to ensure maximum recall a week later? Although much is known about individual factors such as spacing and testing (Brown et al., 2014;

Dunlosky et al., 2013), and there is a growing literature exploring the effect of combining spacing and testing (see Latimier et al., 2021), very little is known about which other combinations of these or other factors are effective and which ones lead to inefficient trade-offs for learners with limited time budgets. Competitions have been successfully employed to address a range of issues in behavioural research, such as identifying successful forecasting methods (Mellers & Tetlock, 2019) and repeated-play strategies (Axelrod, 1984) and hence we set up a competition designed to take a first step towards addressing the question above.

Probably the most important result is that such a competition is feasible. We were able to obtain 13 high-quality learning methods from 11 international teams, and when the 5 best methods were compared in Stage 2 across individuals who completed the task it was possible to distinguish them in terms of their effectiveness as learning methods for individuals self-selected as online language learners. Moreover, all but one of the methods were superior to our *Control* condition, which was very basic in terms of the support it provided for durable learning; the winning solution almost doubled final recall compared to this baseline condition. Other aspects of the results serve to validate the competition's methods: As expected on the basis of decades of research, both age and gender moderated the results.

In a sense the data are not just the retention scores and other measures elicited from the participants, but also the choices made by the teams themselves. Although the competition rules were designed to minimize constraints on the possible solutions, all the Stage 2 solutions employed retrieval practice, suggesting that the groups regarded this as an indispensable feature of any successful method. While this is probably not surprising given the mounting evidence for the benefits of testing, including in real-world classrooms (see Argawal et al., 2021; Trumbo et al., 2021; Yang et al., 2021), other choices are less obvious. Adaptive learning was a popular feature, with three solutions (*Study-Test*, *Errorful Generation* and *Memory Champion*) using adaptive algorithms that ensured new items were presented for study only after previous items had been learned to a

criterion. Three of the methods (*Errorful Generation, Memory Champion,* and *Study-Test*) computed adaptive weights for each word, depending on how well they were remembered and the interval between presentations, to determine how often the word was presented, thus allowing harder words to appear more often. Although adaptive learning is a highly active area in education research (see Xie et al., 2019), it has not been extensively applied in laboratory learning studies. Three of the solutions (*Mediators, Memory Champion*. and *Link Phrases*) instructed participants to generate a mediating keyword or phrase connecting the cue and the target. There is of course a long history of research on the benefits of mnemonics (see Worthen & Hunt, 2010), although the evidence for their effectiveness is somewhat inconsistent (Dunlosky et al., 2013). Overall, while some of the teams' choices involved well-established learning strategies, others reflect innovative ideas regarding optimal methods for foreign-language vocabulary learning. The winning entry, *Memory Champion*, combined a visual mnemonic technique with retrieval practice and an adaptive algorithm for introducing new words. It achieved an average of 27.23, 95% CI [26.08, 28.38] words pairs recalled, so this represents a benchmark against which future methods may be compared, under these learning conditions.

An important aspect of the study is the inclusion of judgments of learning (JOLs) as well as ratings of learning effectiveness and enjoyment. Overall, JOLs and enjoyment ratings aligned to a reasonable extent with actual effectiveness: *Memory Champion* achieved the highest scores on these measures. The fact that a technique is both truly effective and perceived as such (amongst those who completed the study) is significant as it suggests that learners would be motivated to employ the technique in preference to others. On the other hand, and consistent with many similar demonstrations (Rhodes, 2016), there were some clear dissociations between JOLs and retention scores: for instance, while females outperformed males on average recall, their JOLs were equivalent. JOLs also displayed a degree of over-confidence in all solutions except *Mediators,* which was also the solution in which JOLs deviated least from actual scores, suggesting that, of all the solutions, this method gave participants the most accurate metacognitive insight into their own

learning. In terms of actual test scores, however, it was one of the less successful methods. Perhaps the testing of all 80 items at once in the *Mediators* solution made it easier for participants to assess how well they were doing and, particularly, to be aware of how many items they were failing to retrieve, compared with the unbatched presentation of items in the solutions featuring adaptive learning, which were more successful in terms of final test score but which yielded less accurate JOLs. However, *Link Phrases*, which also featured batch testing, did not yield such accurate JOLs, suggesting that a variety of factors may have contributed to participants' metacognitive assessments.

There are of course several limitations of the work described here. By its nature, the competition does not permit us to specify what ingredients differentiate the stronger from the weaker solutions that were entered. We have provided exploratory analyses that shed some light on factors that are associated with success, but these are not definitive because each solution involved a unique combination of features. The competition only evaluates retention after 1 week of 1 hour's learning. It may be that the solutions differ in their effectiveness with longer study periods and retention intervals or with several study periods with different spacing schedules. Table 4 highlights some of the major questions raised by the findings of this competition.

There was considerable attrition in the participants who completed the study. This is probably inevitable in the context of a competition relying on the intrinsic motivation of real learners. Although we measured a range of subjective variables including enjoyment, we have no direct evidence that Memrise learners are more motivated than typical laboratory participants. More importantly, drop-out differed between conditions. As noted above, this itself is an interesting finding, but it does raise the possibility that participants who completed each solution are not equally representative of the population. Only the most motivated participants may have persisted with *Mediators* (which had the lowest completion rate of 21%) whereas participants with a wider range of motivation levels may have persisted with *Study-Test* (36%), thus introducing a potential

confound in the interpretation of the final test scores across solutions (though note that *Memory Champion* outperformed solutions with both lower and higher completion rates). Use of a little-known language, while minimising the possibility that participants would already have encountered the vocabulary, may also have reduced motivation to learn.

The employment of a competition comes with both strengths and weaknesses. While this pragmatic study enabled us to assess effectiveness within the context of Lakota vocabulary learning on Memrise, it does not provide the kinds of theoretical answers that might otherwise be more typical of contemporary research and we have minimal license to generalize its results beyond the language, sample, timing (1 hour of study time, etc.) and so on that we selected. It would be inappropriate to mask these shortcomings.

Despite these and other limitations, we hope that the present work will inspire others to set up comparable competitions. A challenge for the research community is to test the generalizability of, and improve upon, the winning entry, *Memory Champion*.

Acknowledgments

We thank Henry Roediger, Hal Pashler, Elizabeth Maylor, Courtenay Clark, Veronica Yan, and Nicholas Soderstrom for helpful input to the competition rules, Stef Lewandowsky and Eliot York for programming the second phase on behalf of Memrise, and Ji Hae Li and Khuyen Nguyen for assistance with data collection. This work was partly supported by an Impact Acceleration Award from the UK Engineering and Physical Sciences Research Council, EP/K503745/1.

Open Science and Transparency

Data and analysis code are available at https://osf.io/sg6z7/?view_only=a9d00586414d466ab5535545d85d2b75 (Potts & Shanks, 2024). The Supplemental Information includes further details about the competition rules and materials as well as the finalist entries including code for and a video demonstration of the winning method.

Author contributions

References

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review*, *33*, 1409-1453. https://doi.org/10.1007/s10648-021-09595-9

Asperholm, M., Högman, N., Rafi, J., & Herlitz, A. (2019). What did you do yesterday? A meta-analysis of sex differences in episodic memory. *Psychological Bulletin*, *145*, 785-821. https://doi.org/10.1037/bul0000197

Axelrod, R. (1984). *The evolution of cooperation.* Basic Books.

Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014). *Make it stick: The science of successful learning.* Harvard University Press.

Bryfonski, L., & McKay, T. H. (2019). TBLT implementation and evaluation: A meta-analysis. *Language Teaching Research*, *23*(5), 603-632. https://doi.org/10.1177/1362168817744389

Dunlosky, J. & Rawson, K. A. (Eds.) (2019). *The Cambridge handbook of cognition and education*. Cambridge University Press. https://doi.org/10.1017/9781108235631

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*, 4-58. https://doi.org/10.1177/1529100612453266

González-Betancor, S. M., Bolívar-Cruz, A., & Verano-Tacoronte, D. (2019). Self-assessment accuracy in higher education: The influence of gender and performance of university students. *Active Learning in Higher Education*, *20*, 101-114. https://doi.org/10.1177/1469787417735604

Grimaldi, P. J., Pyc, M. A., & Rawson, K. A. (2010). Normative multitrial recall performance, metacognitive judgments, and retrieval latencies for Lithuanian—English paired associates. *Behavior Research Methods*, *42*, 634-642. https://doi.org/10.3758/BRM.42.3.634

Hall, J. F. (1954). Learning as a function of word-frequency. *American Journal of Psychology*, *67*, 138-140.

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology*, *115*, Article 101237. https://doi.org/10.1016/j.cogpsych.2019.101237

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science*, *22*, 787-794. https://doi.org/10.1177/0956797611407929

Latimier, A., Peyre, H., & Ramus, F. (2021). A meta-analytic review of the benefit of spacing out retrieval practice episodes on retention. *Educational Psychology Review*, *33*, 959-987. https://doi.org/10.1007/s10648-020-09572-8

Mellers, B. A., & Tetlock, P. E. (2019). From discipline-centered rivalries to solution-centered science: Producing better probability estimates for policy makers. *American Psychologist*, *74*, 290-300. https://doi.org/10.1037/amp0000429

Morey, R. D., & Rouder, J. N. (2022). BayesFactor: Computation of Bayes factors for common designs. (Version 0.9.12-4.4) [R package]. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Pallier, G. (2003). Gender differences in the self-assessment of accuracy on cognitive tasks. *Sex Roles*, *48*, 265-276. https://doi.org/10.1023/A:1022877405718

Pan, S. C., & Sana, F. (2021). Pretesting versus posttesting: Comparing the pedagogical benefits of errorful generation and retrieval practice. *Journal of Experimental Psychology: Applied*, *27*, 237–257. https://doi.org/10.1037/xap0000345

Porzsolt, F., Rocha, N. G., Toledo-Arruda, A. C., Thomaz, T. G., Moraes, C., Bessa-Guerra, T. R., Leão, M., Migowski, A., Araujo da Silva, A. R., & Weiss, C. (2015). Efficacy and effectiveness trials have different goals, use different tools, and generate different messages. *Pragmatic and Observational Research, 6,* 47–54. https://doi.org/10.2147/POR.S89946

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*, 644-667. https://doi.org/10.1037/a0033194

Potts, R., & Shanks, D. (2024, March 12). *Optimal methods for learning foreign-language vocabulary: An international research competition*. Retrieved from osf.io/sg6z7

Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 65-80). Oxford University Press.

Shortt, M., Tilak, S., Kuznetcova, I., Martens, B., & Akinkuolie, B. (2021). Gamification in mobile-assisted language learning: A systematic review of Duolingo literature from public release of 2012 to early 2020. *Computer Assisted Language Learning, 36*(3), 517–554. https://doi.org/10.1080/09588221.2021.1933540

Trumbo, M. C. S., McDaniel, M. A., Hodge, G. K., Jones, A. P., Matzen, L. E., Kittinger, L. I., Kittinger, R. S., & Clark, V. P. (2021). Is the testing effect ready to be put to work? Evidence from the laboratory to the classroom. *Translational Issues in Psychological Science*, *7*, 332-355. https://doi.org/10.1037/tps0000292

Ward, E. V., Berry, C. J., Shanks, D. R., Moller, P. L., & Czsiser, E. (2020). Aging predicts decline in explicit and implicit memory: A life-span study. *Psychological Science*, *31*, 1071-1083. https://doi.org/10.1177/0956797620927648

Worthen, J. B., & Hunt, R. R. (2010). *Mnemonology: Mnemonics for the 21st century.* Psychology Press.

Xie, H., Chu, H.-C., Hwang, G.-J., & Wang, C.-C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. *Computers & Education*, *140*, Article 103599. https://doi.org/10.1016/j.compedu.2019.103599

Yan, V. X., Eglington, L. G., & Garcia, M. A. (2020). Learning better, learning more: The benefits of expanded retrieval practice. *Journal of Applied Research in Memory and Cognition*, *9*, 204-214. https://doi.org/10.1016/j.jarmac.2020.03.002

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*, 399-435. https://doi.org/10.1037/bul0000309

Appendix A

Stimulus set for Stage 1: Lithuanian-English Word Set

Lithuanian	English	Lithuanian	English
lova	bed	urvas	cave
mesa	meat	augalas	plant
upe	river	zole	grass
sesuo	sister	kede	chair
namas	house	riteris	knight
daina	song	miestas	city
pupa	bean	lietus	rain
akis	eye	sketis	umbrella
nafta	oil	paukstis	bird
karalius	king	ziedas	ring
burna	mouth	raktas	key
gele	flower	puodelis	cup
mokykla	school	adata	needle
tiltas	bridge	bulve	potato
smegenys	brain	sokis	dance
sausainis	cookie	pinigine	wallet
purvas	dirt	palepstis	broom

palaidine	shirt	bugnas	drum
batas	shoe	kunigas	priest
vejas	wind	plaukas	hair
medis	tree	tvora	fence
tinklas	net	arbata	tea
pastatas	building	mygtukas	button
stogas	roof	sakute	fork
pyragas	cake	kreida	chalk
durys	door	vaistas	drug
zirkles	scissors	zuvis	fish
rusys	basement	kumpis	ham
traukinys	train	laiptelis	stair
pienas	milk	laidas	wire
krautuve	store	salmas	helmet
obuolys	apple	smaragdas	emerald
knyga	book	ledas	ice
langas	window	kardas	sword
auksas	gold	kraujas	blood
menulis	moon	plyta	brick
vanduo	water	tvartas	barn

vinis	nail	kablelis	hook
koja	leg	turgus	market
duona	bread	stalas	table

Appendix B

Stimulus Set for Stage 2: Lakota-English Word Set

Lakota	English	Lakota	English
huku	mother	iyeye	find
nape	hand	ble	lake
wicahpi	star	can	tree
mahpiya	cloud	iphiyaka	belt
nako	also	wote	eat
wicasa	man	wicicala	girl
ehate	laugh	naho	hear
ikowayeka	catch	hoksila	boy
tate	wind	tawicu	wife
ge	ask	makoce	earth
iyotake	sit	wata	boat
wiya	woman	cuwitku	daughter
caje	name	mani	walk
we	blood	sota	smoke
ota	more	toha	when
takuwe	why	pasu	nose
ohuta	shore	yuwakol	lift
kimimela	butterfly	wahinkpe	arrow

cakpe	knee	ska	white
tasnaheca	squirrel	ceye	weep
iputake	kiss	suka	dog
wiyaka	feather	magaska	swan
spaya	cook	cepe	fat
mni	water	pilamaye	thanks
tipi	house	hiyu	come
ista	eye	haske	long
hoga	fish	hi	arrive
hanwi	moon	ahco	arm
si	foot	iku	chin
asanpi	milk	taca	body
loci	hungry	kagitaka	raven
sugila	fox	kikta	wake
peta	fire	te	die
hinske	tooth	cante	heart
ohinni	always	htaleha	yesterday
bloketu	summer	wigli	oil
epazo	point	waglula	caterpillar
wakpa	river	heci	there

hoipate	net	zitkala	bird

Table 1

Features of the Five Finalist Solutions and Control Task Implemented in Stage 2 (Memrise Stage)

Solution	Overview of method	Adaptive algorithm?	Keywords instructed or encouraged (as optional strategy)? To cue-target pair or to cue alone?	Initial presentation of word is on its own, before translation appears?
Control	Cue-target pairs presented for study in batches of 80, repeated in a different random order until the hour is up.	No	No use of keywords.	No – every presentation of item is as a pair.
Errorful Generation	Cue alone presented on every trial. Participants respond with a guess on initial presentation, followed by repeated spaced retrieval practice.	Yes	Encouraged, to cue-target pair.	Yes. (Participant to guess meaning before translation appears)
Link Phrases	Cue-target pairs are presented for up to 25 seconds in batches of 10, while participant creates and enters a memorable "link phrase" to link the cue	No	Instructed, to cue-target pair.	No – first presentation of items is

	and the target. Each batch is followed by a multiple choice or matching test. Each set of 40 items is followed by retrieval practice.			always as a pair.
Mediators	All 80 cue-target pairs presented and participant instructed to create mediator between cue and target, followed by repeated retrieval practice in batches of 80.	No	Instructed, to cue-target pair.	No – first presentation of items is always as a pair.
Memory Champion	An introductory video shows a memory champion explaining how to generate a memorable mental image from a cue, which is presented alone, and associate that image with the subsequently-presented target. Items are encoded against images of rooms and participants are encouraged to associate their mediating images with the corresponding room, followed by repeated spaced retrieval practice.	Yes	Instructed (but not required to be typed), to cue alone.	Yes. (Participant to generate keyword + image before translation appears)
Study-Test	Each cue-target pair presented for study once, followed by repeated spaced retrieval practice, with corrective	Yes	No use of keywords.	No – first presentation of items is

feedback displayed only when the response was incorrect.	always as a pair.

Table 2

Progress of Participants in Stage 2

Solution	Invited (i.e., sent the link to the study phase)	Started study (% of those invited in brackets)	Completed study (% of those who started) – *all those who have a JOL.*	Completed test (% of those who started the experiment) – *all those with final test score.*	Completed w/in rules and within permitted retention interval (% of those who started) *With final test score, and "1" in the three honesty columns, first time = true, retention interval 6-9 days*)
Control	6981	2150 (31%)	925 (43%)	784 (36%)	535 (25%)
Errorful Generation	6900	2153 (31%)	1050 (49%)	950 (44%)	692 (32%)
Link Phrases	6809	2173 (32%)	805 (37%)	675 (31%)	459 (22%)

Mediators	6805	2075 (30%)	753 (36%)	630 (30%)	475 (21%)
Memory Champion	9226[2]	2771 (30%)	1331 (48%)	1167 (42%)	863 (31%)
Study-Test	6933	2151 (31%)	1164 (54%)	1037 (48%)	779 (36%)
Totals	43654	13473 (31%)	6028 (45%)	5243 (39%)	3803 (28%)

Note. Table 2 shows the progress of participants during Stage 2 for each solution, from signing up to completing the experiment.

[2] Due to a software error, there was a brief period when the probability of participants being allocated to Memory Champion was 2 in 7 rather than 1 in 6, leading to higher recruitment to that solution overall.

Table 3

Participants' Metacognitive Ratings in Stage 2

Solution	Number of participants in final dataset	JOL (1 – 80)	Enjoyment rating at end of study (1 - 5)	Enjoyment rating at end of test (1 - 5)	Effectiveness rating at end of study (1 – 5)	Effectiveness rating at end of test (1 – 5)
Control	535	27.92 (20.31)	2.67 (1.13)	2.80 (1.17)	2.41 (.96)	1.78 (.87)
Errorful Generation	692	28.51 (22.06)	3.25 (1.20)	3.31 (1.17)	2.97 (1.09)	2.45 (1.07)
Link Phrases	459	27.54 (18.57)	3.25 (1.03)	3.09 (1.09)	3.21 (.95)	2.30 (.97)
Mediators	475	18.54 (15.84)	3.10 (1.15)	3.37 (1.33)	3.04 (1.00)	2.72 (1.04)
Memory Champion	863	39.20 (20.73)	3.83 (.94)	3.65 (.99)	3.69 (.91)	2.90 (.99)
Study Test	779	31.20 (20.98)	3.39 (1.10)	3.42 (1.11)	3.09 (.98)	2.51 (1.00)

Note. Sample size, mean judgments of learning (JOL), and enjoyment and effectiveness ratings for each solution in Stage 2. Numbers in brackets are *SD*s.

Table 4

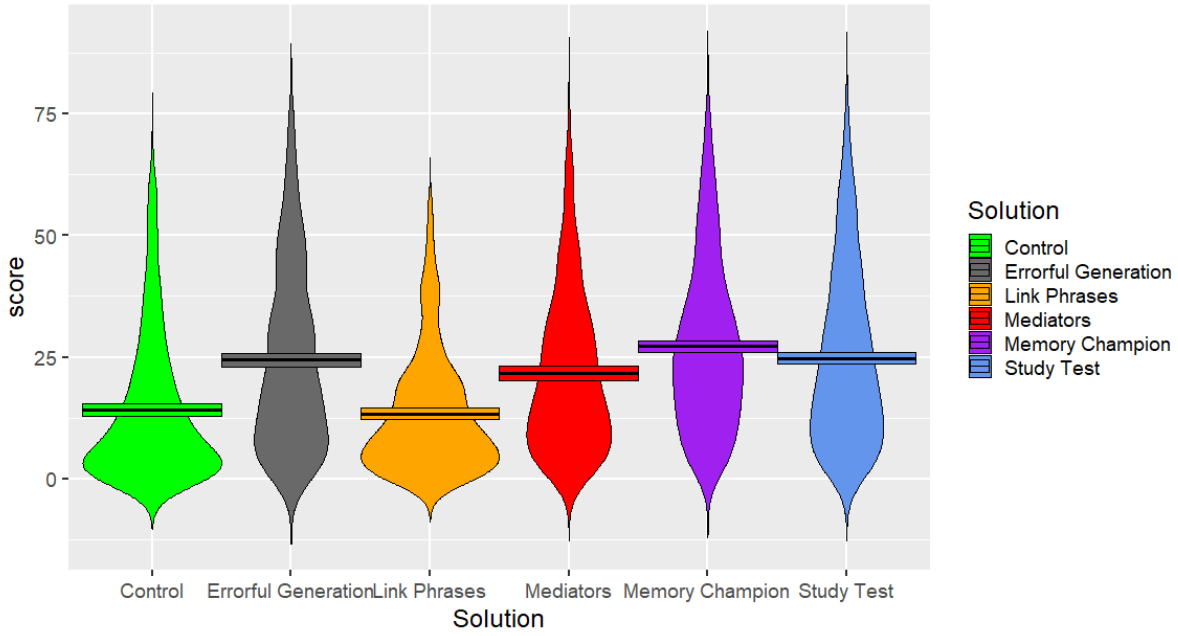
Key Outstanding Questions from the Competition about Optimizing Learning in Real World Contexts

	Questions for future research
1	Would the ordering of the finalist solutions be the same with different languages and/or with different word sets, featuring different parts of speech?
2	Would the ordering of the finalist solutions be the same with different amounts of initial study time and different recall delays?
3	Is the winning method (*Memory Champion*) best for learning different kinds of information, such as historical dates, medical facts, etc.?
4	Would the relative ordering of the solutions be different if learning were distributed over several shorter phases rather than a single 1-h session and/or over multiple spaced sessions on different days?
5	What individual difference variables (if any) determine the relative effectiveness of each solution for a given person?
6	Does practice in applying a given solution enhance its effectiveness?
7	What is the most effective number of retrieval practice trials and how might this vary with different spacing intervals?
8	What features of the winning method (*Memory Champion*) are essential for its success? Are these features additive or interactive in their influence?
9	How could the winning method (*Memory Champion*) be improved upon?

10 What features of the finalist solutions have most influence on enjoyment ratings and how might these ratings change as learning continues over longer time periods (weeks and months)?

Figure 1

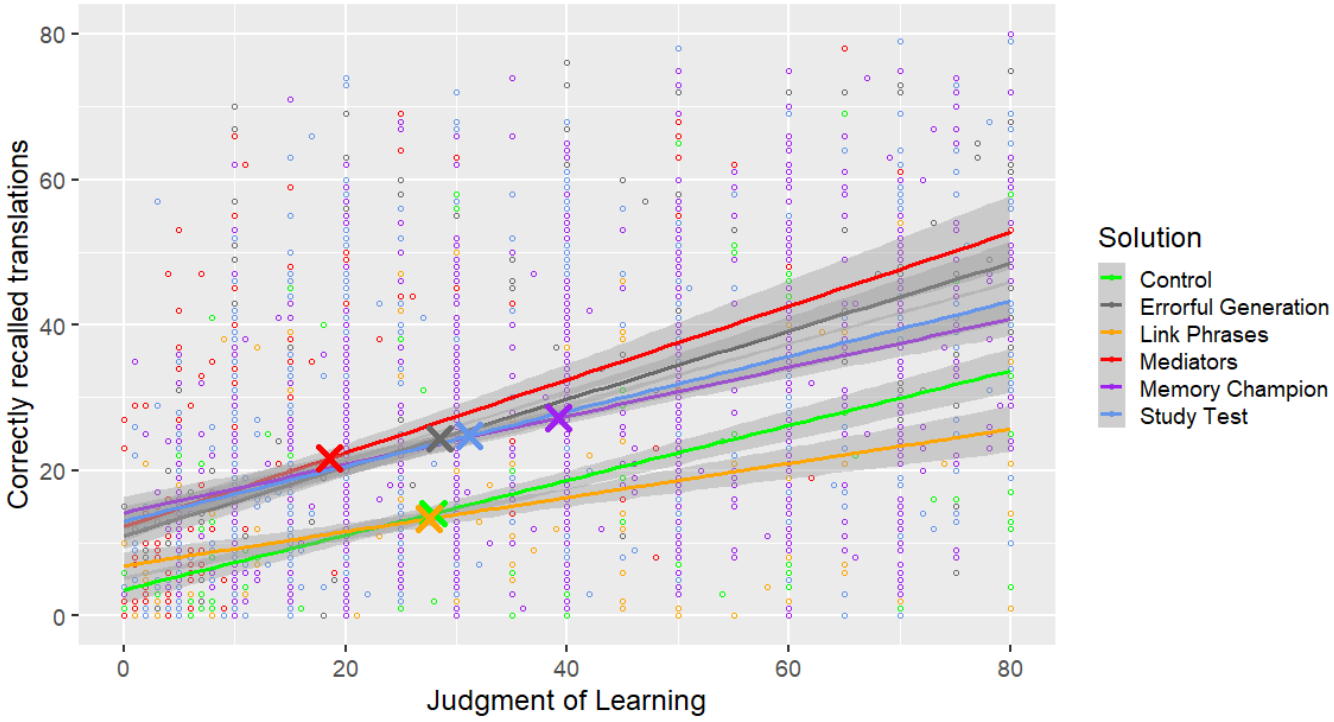
Final Test Scores (out of 80) for each Method in Stage 2 (Memrise Stage).



Note: The plot shows the mean (black bars), 95% confidence interval (coloured bars), and smoothed densities for each method.

Figure 2

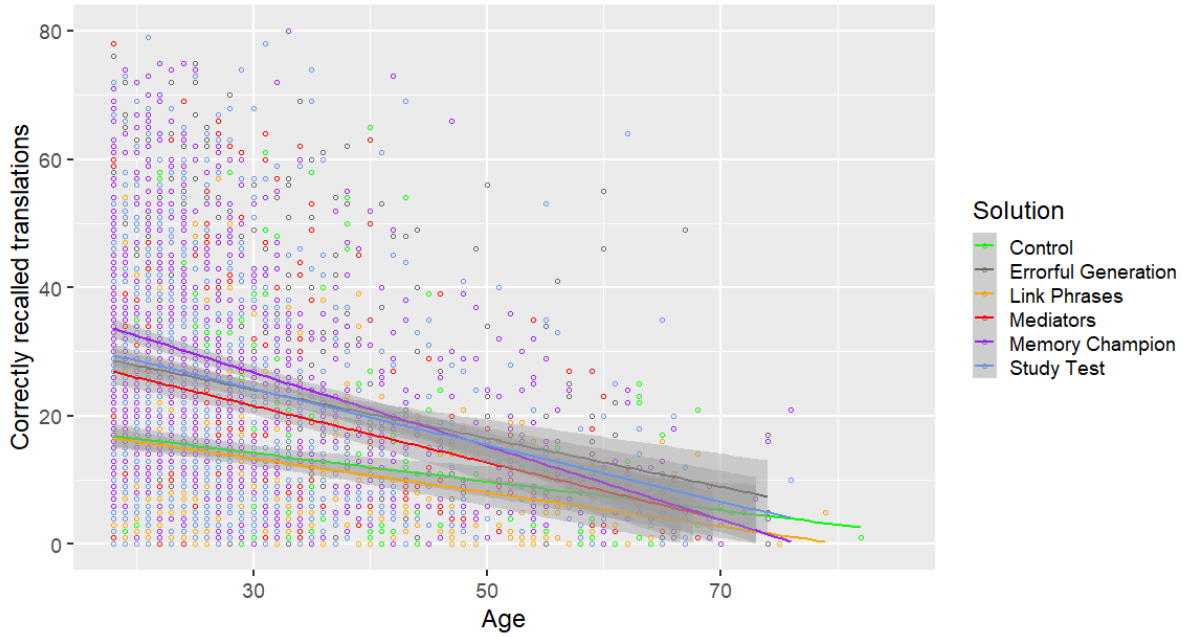
JOLs by Final Score for Each Solution in Stage 2, With Best-fitting Linear Regression Lines.



Note. The grey regions are 95% confidence intervals. Crosses show solution means.

Figure 3

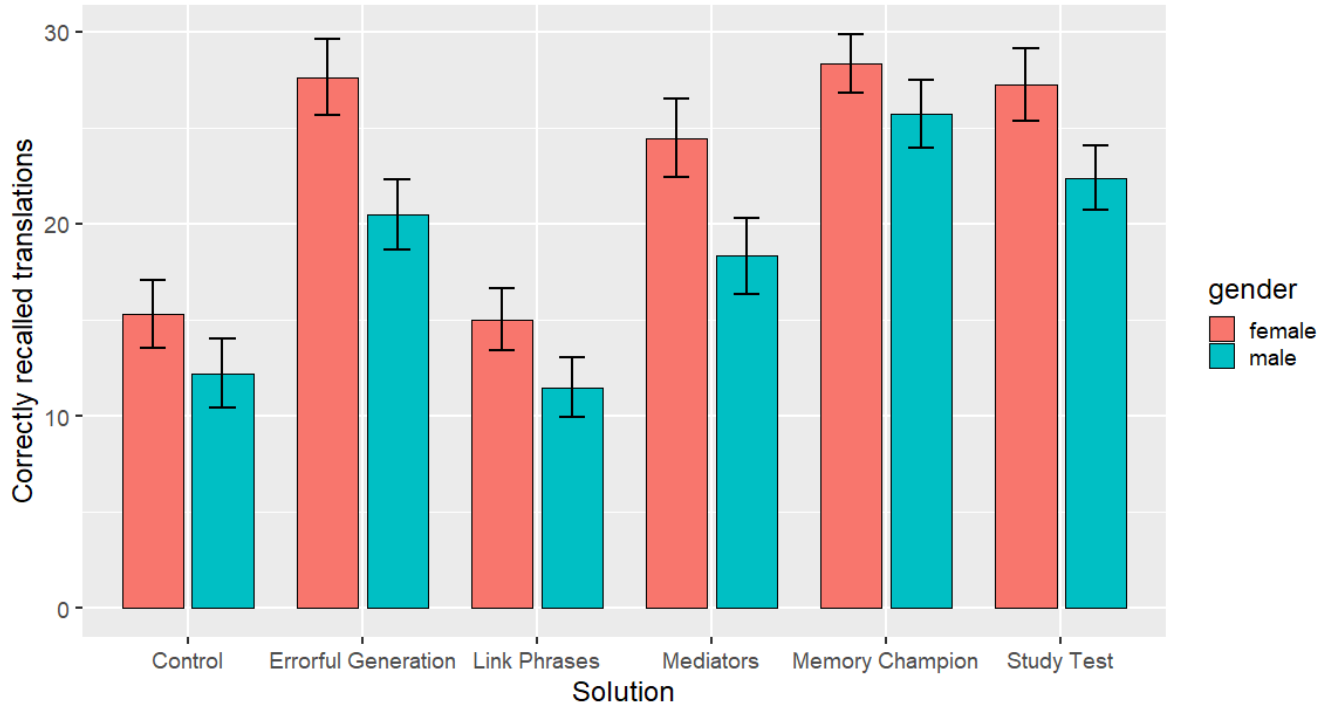
Final Test Scores by Age for all Participants in Stage 2.



Note. The figure shows final test scores by age for all participants, split by solution, with best-fitting linear regression lines, in Stage 2. The grey regions are 95% confidence intervals.

Figure 4

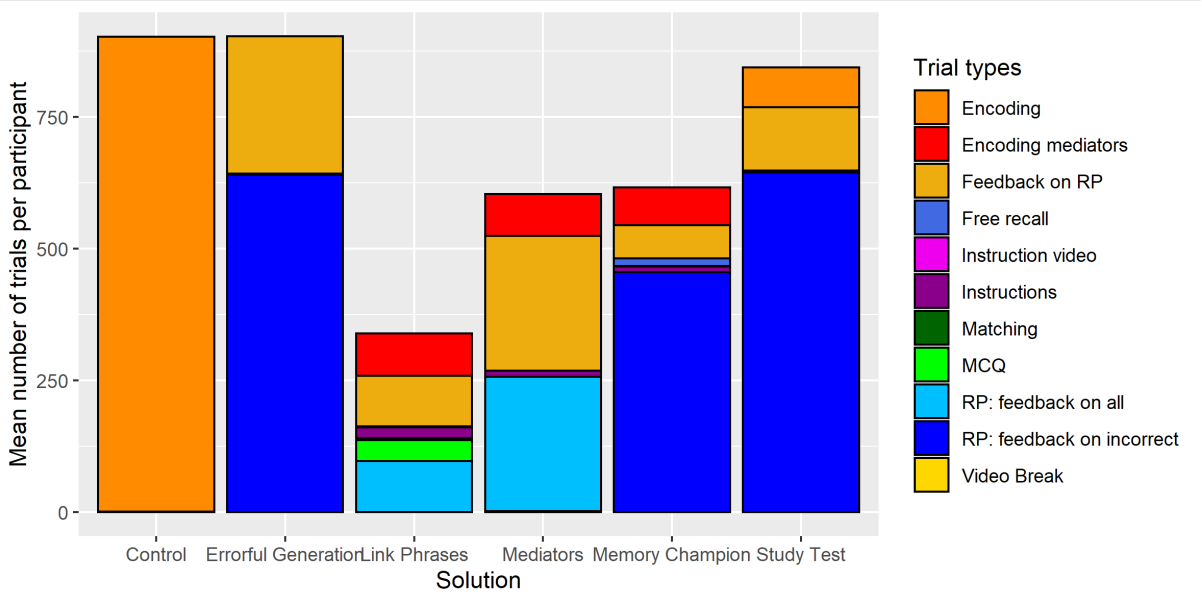
Final Test Score by Gender in Stage 2



Note. Final test score by gender for each solution in Stage 2 (error bars are 95% confidence intervals).

Figure 5

Breakdown of Trial Types by Solution in Stage 2.



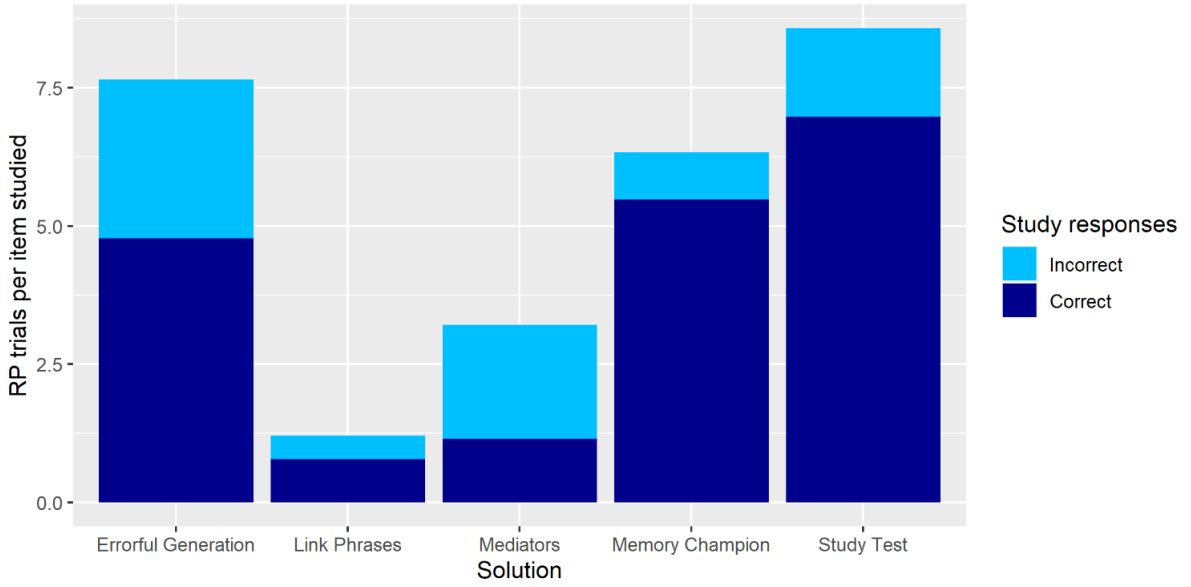
Note. Key to trial types: *Encoding*: a pure study trial, i.e., a trial on which a cue-target pair is presented, other than when it is presented as feedback following a retrieval practice trial. *Encoding mediators*: a trial on which the participant is to produce a mediator or keyword and either type it in or simply think it. *Feedback on RP:* A cue-target presentation that follows a retrieval practice trial. *Free recall*: Applies only to the *Memory Champion* method and represents trials on which an image of a room is presented and the participant is encouraged to think of all the items that were studied against the backdrop of that image (covert retrieval). *Instruction video*: A video explaining how to study the vocabulary pairs. *Instructions*: written instructions. *Matching*: trials on which both Lakota and English words are presented and the participant is to match the correct pairs. *MCQ*: trials on which a cue is presented with a choice of possible English translations and the participant is to select the correct translation. *RP: feedback on all*: trials on which cues are presented for retrieval practice, following which the correct cue-target pair is presented whether or not the participant responded correctly. *RP: feedback on incorrect*: trials on which cues are presented for retrieval practice,

following which the correct cue-target pair is presented only when the participant has responded incorrectly (though correct answers may have been confirmed by a brief visual or auditory signal).

Video break: Applies only to the *Mediators* method and represents the showing of a one-minute video of a waterfall.

Figure 6

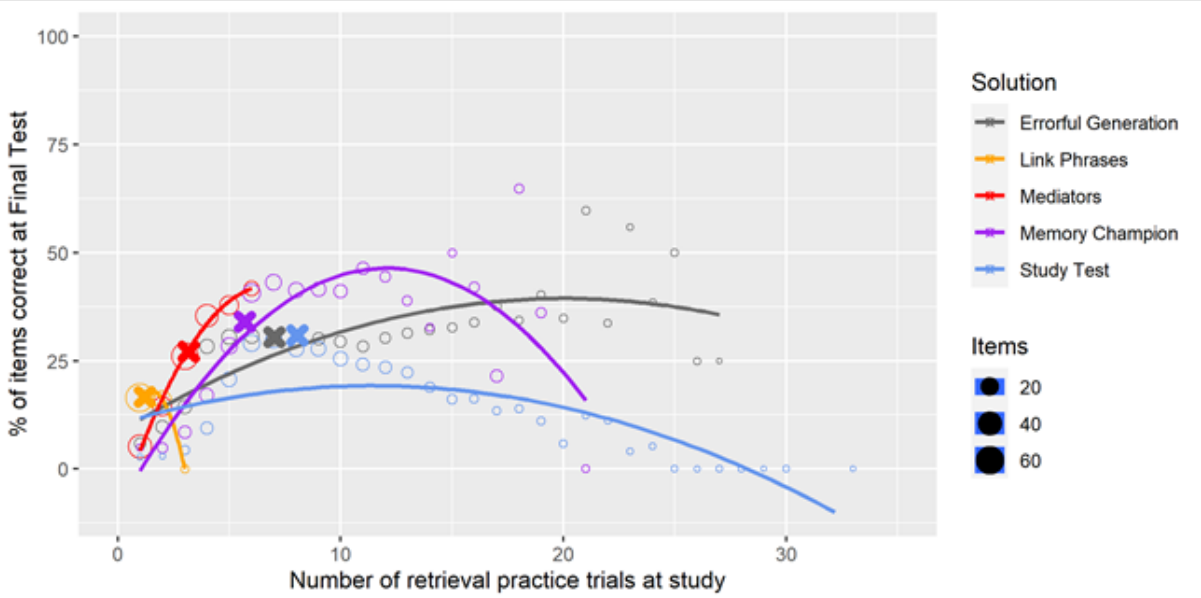
Retrieval Practice Trials, Correct or Incorrect at study, in Stage 2.



Note. Figure 6 shows the mean number of retrieval practice trials per studied item, split by whether the response was correct or incorrect at study, in Stage 2.

Figure 7

Final Test Scores as a Function of the Number of Retrieval Practice Trials in Stage 2.



Note. The figure shows the percentage of items correctly recalled at final test as a function of the number of retrieval practice trials for those items at study.The sizes of the datapoints are proportional to the number of observations. The crosses indicate, for each condition, the mean final test score and mean number of retrieval practice trials. Note that for clarity of presentation, the plot does not include instances where, across all participants in a solution, there were fewer than 3 items practiced at a given number of retrieval practice trials. This results in the omission of 12 datapoints (3 x Errorful Generation, 2 x Memory Champion, 7 x Study Test) where, in each case, either one or two items were practiced between 20 and 51 times.