The Journal of Experimental Education

Routledge
Taylor & Francis Group

ISSN: 0022-0973 (Print) 1940-0683 (Online) Journal homepage: https://www.tandfonline.com/loi/vjxe20

Retrieval Practice Enhances Analogical Problem Solving

Sarah Shi Hui Wong, Gavin Jun Peng Ng, Tobias Tempel & Stephen Wee Hun Lim

To cite this article: Sarah Shi Hui Wong, Gavin Jun Peng Ng, Tobias Tempel & Stephen Wee Hun Lim (2019) Retrieval Practice Enhances Analogical Problem Solving, The Journal of Experimental Education, 87:1, 128-138, DOI: 10.1080/00220973.2017.1409185

To link to this article: https://doi.org/10.1080/00220973.2017.1409185

Published online: 26 Dec 2017.

Submit your article to this journal

Article views: 169

View Crossmark data

Citing articles: 1 View citing articles

THE JOURNAL OF EXPERIMENTAL EDUCATION
2019, VOL. 87, NO. 1, 128–138
https://doi.org/10.1080/00220973.2017.1409185

Retrieval Practice Enhances Analogical Problem Solving

Sarah Shi Hui Wong[a], Gavin Jun Peng Ng[a], Tobias Tempel[b], and Stephen Wee Hun Lim[a]

[a]Department of Psychology, National University of Singapore, Singapore; [b]Fachbereich I – Psychologie, University of Trier, Trier, Germany

ABSTRACT

The impact of retrieval practice on analogical-problem-solving performance was investigated using a complex, educationally relevant task. Participants studied a statistical hypothesis testing scenario and practiced recalling the material or repeatedly studied it. Participants then completed a final test either 5 minutes or 1 week later involving a novel hypothesis-testing scenario that shared an intermediate procedural strategy and superficial and structural similarity with the study scenario but that differed at a specific procedure level. When the final test was given after 5 minutes, no differences in performance were observed across conditions ($d = 0.01$). Crucially, on the delayed test, retrieval practice produced superior performance than did repeated studying ($d = 0.81$), whereby participants were better at applying learned knowledge to solve a novel problem.

KEYWORDS

Analogical problem solving; procedural knowledge; retrieval practice; testing; transfer of learning

THE APPLICATION OF prior learning to solve novel problems is a fundamental goal of education. Yet, while transfer of learning has been heralded as "the ultimate aim of teaching," it is also considered "one of teaching's most formidable problems" (McKeough, Lupart, & Marini, 1995, p. vii). To address the challenge of facilitating transfer of learning, we investigated and found support for the use of retrieval practice as a promising approach to enhance analogical problem solving. Specifically, learners who had practiced recalling a statistical hypothesis-testing scenario subsequently performed better than those who had repeatedly studied it on a delayed final test requiring the flexible application of the learned analogous solution to solve a novel hypothesis-testing problem. This finding poses pedagogical implications for promoting learners' retention and transfer of procedural knowledge through integrating retrieval practice into educational activities.

Analogical problem solving

When we encounter a new problem, we tend to make reference to similar problems that we have dealt with previously, adopting and applying past solution methods. This is known as analogical problem solving and involves the transfer of previously acquired knowledge from one situation to another. Analogical problem solving entails three separate processes—the retrieval of learned (source) problems and their solutions, the abstraction of the underlying solution principle, and the application of this knowledge to solve novel (target) problems when the analogical relations between them have been noticed and mapped. According to this view, retrieval is a prerequisite for knowledge application and problem solving. Once the knowledge is successfully retrieved, the extent of transfer then hinges largely on the

CONTACT Stephen Wee Hun Lim psylimwh@nus.edu.sg National University of Singapore, Faculty of Arts & Social Sciences, Department of Psychology, Block AS4, 9 Arts Link, 117570, Singapore.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/vjxe.

nature and degree of task similarity between the source and target problems. Specifically, there are three commonly identified types of task similarity: surface, structural, and procedural similarity (Chen, 2002; Gentner, Ratterman, & Forbus, 1993; Gick & Holyoak, 1983). While *surface similarity* refers to common peripheral (solution-irrelevant) information on, for instance, objects or characters contained in source and target problems, *structural similarity* relates to the preservation of causal relations among the key components of the goal structure shared by the source and target problems, including matches in goals, obstacles, resources, and a solution principle to attain an outcome. Relatedly, *procedural similarity* entails a match between the solution in the source analog and the required target solution. The extent of procedural similarity may occur at three distinct levels in order of increasing concreteness of the contexts shared: a superordinate principle (abstract general solution) level, an intermediate strategy level, or a specific procedure (concrete operational details) level. Research has established that transfer of knowledge is improved when both the source and target problems are analogous (structurally and procedurally similar) such that the transfer distance is minimized, thereby enabling correspondences between the source and target to be more fruitfully mapped and promoting the direct and straightforward execution of relevant procedures (e.g., Chen, 2002; Holyoak & Koh, 1987). The question we asked was whether, and to what extent, a target problem that is analogous to the source problem could be solved effectively via the use of retrieval practice as a learning strategy.

Retrieval practice

A burgeoning body of research suggests that retrieval practice can promote meaningful learning (e.g., Chan & McDermott, 2007; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Fiorella & Mayer, 2015, 2016; Roediger & Butler, 2011). As opposed to repeatedly studying educational materials, alternating between studying and retrieving the materials from memory (e.g., through free recall tests) has consistently produced superior long-term retention on delayed tests given a week later (e.g., Gates, 1917; Roediger & Karpicke, 2006). This benefit has been demonstrated not only in laboratory settings (e.g., Karpicke & Blunt, 2011; Karpicke & Roediger, 2008; Lim, Ng, & Wong, 2015; Yong & Lim, 2016) but also in real-world school settings (e.g., Carpenter, Pashler, & Cepeda, 2009; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Roediger, Agarwal, McDaniel, & McDermott, 2011). A growing number of studies (for a review, see Karpicke & Aue, 2015) have provided evidence of retrieval practice effects using complex materials (see, however, van Gog & Sweller, 2015), including passages describing the operation of mechanical devices such as brakes and pumps (McDaniel, Howard, & Einstein, 2009) and texts with sequence structures describing a continuous and ordered series of events such as digestion processes (Karpicke & Blunt, 2011). Besides declarative knowledge, testing also benefits the retention of procedural knowledge, including motor sequences (Tempel & Kubik, 2017) and resuscitation skills (Kromann, Jensen, & Ringsted, 2009). Furthermore, retrieval practice promotes the transfer of factual and conceptual knowledge to tests that require inferences to be drawn from the previously learned information (e.g., Butler, 2010; Chan, McDermott, & Roediger, 2006). Taken together, it is logical to hypothesize that the benefits of retrieval practice extend to complex materials in analogical problem solving, which requires the procedural knowledge and step-wise application of specific rules to execute an algorithm to solve a problem (Star, 2005).

The present study

In our study, we employed a relationally complex passage in the domain of statistical hypothesis testing that required learners to make connections among a series of interrelated steps to fully understand the problem-solving process and algorithm (e.g., van Gog & Sweller, 2015). The target problem we constructed shared the same knowledge domain as the source problem on statistical hypothesis testing and was also superficially and structurally similar. Crucially, in terms of procedural similarity, the target shared an intermediate strategy with the source but differed at a specific procedure level. Thus, successful transfer demanded that learners not only recall the relevant solution principle, but also apply it to implement similar yet nonidentical concrete procedures. This allowed us to examine the effect of retrieval practice on an educationally relevant and relatively complex task, going beyond Duncker's (1945) radiation problem that has been commonly used in analogical-problem-solving research. We expected retrieval practice using free recall to improve access to the algorithm, thereby enhancing analogical-problem-solving performance relative to repeated studying, particularly in a delayed final test.

Method

Participants

Sixty undergraduate students (31 were female) between the ages of 18 and 26 from the National University of Singapore (NUS) took part in the study and were each reimbursed $10 for an hour of participation. The students came from the following faculties of study: Arts and Social Sciences (18), Business (6), Computing (5), Engineering (18), Medicine (1), and Science (12). Only students who declared that they had no prior knowledge in statistics and/or had not read statistics modules at the NUS were recruited via an advertisement mounted through the university's online Research Participation Programme SONA system and allowed to participate. This research was conducted with the appropriate ethics-review-board approval by the NUS, and participants granted their written informed consent.

Design

A 2 × 2 × 2 fully between-subjects design was used. The two main independent variables (IVs) of interest were (a) learning condition (S denotes study; R denotes retrieval practice): repeated study (SSSS) versus study interleaved with retrieval practice (SRSR), and (b) retention interval: 5-minute versus 1-week retention interval. A third IV, study passage—"sprint" versus "sleep"—was included purely for control purposes (i.e., to ensure that effects, if any, persisted across passage topics). The dependent variable was analogical-problem-solving performance as operationalized by the mean proportion of idea units correctly recalled and applied in a final test.

Materials

Two prose passages were used (please see Appendices A and B). The passages comprised different scenarios involving a specific hypothesis to be tested: "Does drinking coffee enable sportsmen to sprint faster?" ("sprint") versus "Does drinking milk before one sleeps enhance sleep quality?" ("sleep"), but were fundamentally equivalent in their underlying algorithm comprising statistical (hypothesis testing) concepts and procedures. For example, both scenarios involved restating the question as a research (alternative) hypothesis and a null hypothesis about the population, determining the characteristics of the comparison population, and establishing the extreme cutoff score on the comparison population, beyond which the research hypothesis should be accepted. The "sprint" passage contained 671 words and the "sleep" passage contained 776 words; Each passage was decomposable into 34 idea units for scoring purposes.

Procedure

Participants underwent two experimental phases in groups of four or fewer. Phase 1 was a learning phase that consisted of four consecutive periods, each spanning 9 minutes. Participants were informed in advance about taking a final test that would be topically different from, but conceptually similar to, the study material and to bear this in mind while studying. This ensured that all participants would notice the analogous relationship between the study and test scenarios. Participants were randomly assigned to populate either the SSSS (29 participants) or the SRSR (31 participants) condition. Within each learning condition, half of the participants were randomly allocated to study the "sprint" passage whereas the other half of them studied the "sleep" passage. Participants in the SSSS condition studied the material for four consecutive periods. During each study period, participants were instructed to study once through and, thereafter, revisit parts of which they found to be more difficult until the time allocated had expired. Participants in the SRSR condition interleaved studying with retrieval practice. During each retrieval period, participants were instructed to recall freely and write down as much of the material as they could remember. After each of the four periods, all participants completed a filler task (solved math questions). This timed filler task spanned 1 minute after the first three periods, and 5 minutes after the fourth period.

At the end of Phase 1, an interim questionnaire was administered, in which participants indicated on a 7-point Likert scale: (a) how interesting they thought the study material was (1 = *very boring*; 7 = *very interesting*); (b) how understandable the material was (1 = *very difficult to understand*; 7 = *very easy to understand*); (c) how well they thought they would remember the material after 5 minutes or after 1 week, depending on which condition the participants were in (1 = *not very well*; 7 = *very well*); and (d) how well they knew the subject matter covered in the material (1 = *not very well*; 7 = *very well*).

In Phase 2, participants undertook a final test either 5 minutes or 1 week after the learning phase. In the 5-minute-retention-interval condition, the aforementioned task of solving math questions served as the filler task after the fourth period. During the final test, participants were presented with a novel hypothesis-testing scenario—if they had earlier studied the "sprint" passage, they now received the "sleep" scenario, and vice versa—and were asked to write down how they would test the given hypothesis (please see Appendix C). Importantly, the test scenario contained procedural details that differed from what participants had earlier studied during the learning phase, preventing any regurgitation and use of verbatim knowledge per se. For instance, the research hypothesis in a study (e.g., "sprint") passage ($\mu_1 < \mu_2$, where Population 1 drank coffee whereas Population 2 did not, with μ measuring average sprinting time in seconds) contrasted with that required in a test ("sleep") passage ($\mu_1 > \mu_2$, where Population 1 drank milk whereas Population 2 did not, with μ measuring average sleep quality and higher numerical ratings representing better quality). Participants were reminded that, while the test scenario differed contextually from the scenario they had studied in Phase 1, the same statistical concepts and procedures undergirded both scenarios and that the participants were to recall and apply what they had earlier studied to solve the test scenario. This final test lasted 10 minutes, following which, participants were debriefed and reimbursed for their participation.

Results

Two raters independently scored all participants' responses on the final test by awarding one point for each idea unit that had been correctly recalled and applied within the context of the novel test scenario. For instance, for the "sprint" scenario, each of the following four idea units was awarded one point each: "Population 1: Sportsmen who drank coffee," "Population 2: Sportsmen who didn't drink coffee," "Research hypothesis: $\mu_1 < \mu_2$," and "Null hypothesis: $\mu_1 \geq \mu_2$". Inaccurate responses (e.g., "Research hypothesis: Drinking coffee does not enable sportsmen to sprint faster"; "Null hypothesis: Drinking coffee affects the speed of sportsmen by enabling them to sprint faster") were not awarded any points. A third rater resolved discrepancies to reach 100% agreement. The proportion of idea units recalled was computed for each participant by dividing their recall score by the maximum score of 34. These proportion scores were then used in subsequent analyses. One outlying data point falling beyond 2.5 standard deviations from the group mean was identified and excluded, leaving a final sample of 59 participants.

The data were submitted to a 2 × 2 × 2 analysis of variance (ANOVA) with learning condition (SSSS or SRSR) and retention interval (5 minutes or 1 week) as the independent variables of interest and study passage ("sprint" or "sleep") as an independent variable purely for control purposes. The distribution of participants across the critical cells was as follows: SRSR (5 minutes) = 16; SSSS (5 minutes) = 13; SRSR (1 week) = 15; SSSS (1 week) = 15, respectively. As expected, the 3-way interaction did not reach significance, $F < 1$, while a significant learning condition × retention interval interaction emerged, $F(1, 51) = 4.55$, $MSe = 0.02$, $p = .038$, $\eta_p^2 = .082$. To illuminate the specific pattern of results, post hoc analyses were conducted. Whereas no differences were observed in test performance across the SSSS ($M = .40$, $SD = .12$) and SRSR ($M = .40$, $SD = .11$) conditions when the final test was administered immediately, $F < 1$, $d = 0.01$, the crucial finding was that participants in the SRSR condition performed significantly better (M = .29, SD = .17) than those in the SSSS condition (M = .15, SD = .11) when the test was delayed until a week later, $F(1, 51) = 8.27$, $MSe = 0.02$, $p = .006$, $\eta_p^2 = .14$, $d = 0.81$.

To ascertain that the effects observed were not attributable to differences in participants' perceptions of how interesting or understandable the study material was, their metacognitive judgments of learning, and their prior knowledge of the material, a two-factor (learning condition and retention interval) multivariate analysis of variance (MANOVA) was performed with participants' ratings on the four questionnaire items as the dependent variables. The MANOVA indicated that there was no significant learning condition × retention interval interaction across all four questionnaire items, $F < 1$.

Discussion

Our study revealed that analogical-problem-solving performance in a delayed test is enhanced by retrieval practice, relative to repeated studying. This advantage did not emerge when learners were tested immediately, echoing the robust findings in the retrieval-based-learning literature (e.g., Roediger & Karpicke, 2006). The interpretation is that, whereas repeated studying leads to poorer conceptual organization due to more item-specific processing with increased exposure time, repeated retrieval leads to more robust organization that produces more stable recall over time (Congleton & Rajaram, 2011, 2012). While greater item-specific processing during repeated studying may be useful for immediate tests that solely involve the recall of information (e.g., Roediger & Karpicke, 2006), this advantage may well be outweighed by its associated cost of reduced relational processing, particularly in analogical problem solving that demands not only the successful recall of source knowledge but also the appropriate application of this knowledge. This may account for the similar performance eventually observed on the immediate final test across the repeated studying and retrieval practice conditions in our study. As repeated studying produces relatively tenuous organization, however, it does not buffer against memory decay over time as compared to the stronger organization associated with repeated retrieval (Congleton & Rajaram, 2012), thus explaining our finding that participants who had engaged in retrieval practice outperformed those who had studied repeatedly on the delayed final test.

The present research contributes to the recent debate on the extent to which the testing effect holds for complex educational materials (e.g., Karpicke & Aue, 2015; van Gog & Sweller, 2015). Here, we show that the advantages of retrieval practice apply even to a passage on statistical hypothesis testing that contains high element interactivity with various interrelated information elements—or what van Gog and Sweller (2015) define as "complex" material. While Lim et al. (2015) have reported that retrieval practice boosts the retention of verbatim knowledge in statistical hypothesis testing, we further show that these benefits extend to deep comprehension in applying procedural knowledge to solve a similar but nonidentical problem even after a week following initial learning.

Educational implications

It is worth noting that retrieval practice led to superior learning outcomes on an educationally relevant task in our study even though participants had not received prior training on this technique, suggesting that retrieval practice can be readily implemented in the classroom to enhance analogical problem solving. Spontaneous analogical transfer is often fraught with difficulty for learners (e.g., Salomon & Perkins, 1989). As such, educators can guide students to incorporate retrieval practice into their learning routines (e.g., encourage self-testing) to aid durable retention and transfer of knowledge across novel contexts, thereby enhancing analogical problem solving. Educators could, also, actively implement retrieval practice in the classroom—for example, administer low-stakes quizzes (Roediger & Pyc, 2012) that test learners' core domain knowledge and competencies (i.e., present learners with opportunities to retrieve and remember what they have been taught) and provide feedback after the quizzing to amplify the benefits of retrieval (Butler & Roediger, 2008) before subjecting learners to various problem-solving scenarios.

Future directions

In our study, learners engaged in four consecutive learning periods (SSSS versus SRSR), in line with the retrieval-based learning paradigm typically used to investigate the effects of repeated studying versus repeated retrieval (e.g., Karpicke & Blunt, 2011; Roediger & Karpicke, 2006). It is plausible, however, that learners may have found the repetitive nature of the task tedious and boring, such that increased distractibility and mind wandering occurred over the four periods, in turn lowering overall absolute performance on the final test. To address this potential problem, future work may, as suggested by an anonymous reviewer of our work, consider examining the effects of a single retrieval attempt (i.e., SS versus SR) on analogical problem solving and with larger sample sizes to ensure that the studies are statistically well-powered.

At the same time, the final test in our study involved a written essay in which participants explained how they would solve a novel hypothesis-testing problem by implementing a previously learned solution procedure. Future studies may alternatively assess analogical problem solving in the domain of statistical hypothesis testing through requiring learners to, for example, test a novel hypothesis by analyzing a given data set to solve for the actual t statistic and p value. In addition, the target problem we designed was analogous to the source problem superficially, structurally, and procedurally at an intermediate strategy level within the same knowledge domain—conditions that facilitate near transfer of a learned solution procedure. Future research can investigate whether retrieval practice may aid far transfer of procedural knowledge by varying dimensions such as test format, physical context, and knowledge domain (Barnett & Ceci, 2002). Indeed, there is evidence that retrieval practice produces better transfer of facts and concepts even to questions in different knowledge domains (Butler, 2010; see also Carpenter, 2012, for a review), suggesting that the advantages conferred by repeated retrieval are sufficiently robust to overcome contextual variation.

Furthermore, analogical transfer within knowledge domains such as mathematics may be difficult for many students even when the source and target problems are highly analogous in the solution procedures that they require (e.g., Reed, Dempster, & Ettinger, 1985). While extant studies have examined scaffolding methods that make mathematical problem structures more transparent to learners in order to facilitate transfer (e.g., Lee, Betts, & Anderson, 2017), considerably less attention has been devoted to investigating techniques that promote the long-term retention of learned procedures necessary for transfer to occur when the source and target problems have been successfully mapped. Accordingly, it may be educationally useful for future work to explore whether retrieval practice facilitates analogical transfer of procedural knowledge across other domains such as mathematical problem solving, beyond our study's focus on statistical hypothesis testing.

Conclusion

Transfer of knowledge is a fundamental aim of education—the capacity to apply knowledge to markedly different situations allows us to make sense of an unpredictable world, learn new skills efficiently, and creatively solve novel problems (Haskell, 2001). Given the premium placed on transfer in learning, promoting the conditions and techniques that facilitate transfer is crucial. Using relatively complex and educationally relevant materials, the present study has revealed that retrieval practice is one promising technique that enhances analogical problem solving. Much potential is embodied in implementing retrieval practice as an effective tool to foster deep, meaningful learning.

Funding

The research was supported in part by a National University of Singapore Faculty of Arts and Social Sciences Heads & Deanery Research Support Scheme grant (R-581-000-160-101) awarded to LimSWH.

References

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, *128*, 612–637. doi:10.1037/0033-2909.128.4.612

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1118–1133.

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

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

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, *23*, 760–771. doi:10.1002/acp.1507

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

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*, 553–571. doi:10.1037/0096-3445.135.4.553

Chen, Z. (2002). Analogical problem solving: A hierarchical analysis of procedural similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 81–98.

Congleton, A., & Rajaram, S. (2011). The influence of learning methods on collaboration: Prior repeated retrieval enhances retrieval organization, abolishes collaborative inhibition, and promotes post-collaborative memory. *Journal of Experimental Psychology: General*, *140*, 535–551. doi:10.1037/a0024308

Congleton, A., & Rajaram, S. (2012). The origin of the interaction between learning method and delay in the testing effect: The roles of processing and conceptual retrieval organization. *Memory and Cognition*, *40*, 528–539. doi:10.3758/s13421-011-0168-y

Duncker, K. (1945). On problem-solving. *Psychological Monographs*, *58*(5), i–113 (Whole No. 270). doi:10.1037/h0093599

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.

Fiorella, L., & Mayer, R. E. (2015). *Learning as a generative activity: Eight learning strategies that promote understanding*. New York: Cambridge University Press.

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*, 717–741.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *6*(40), 1–104.

Gentner, D., Ratterman, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, *25*, 524–575. doi:10.1006/cogp.1993.1013

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38. doi:10.1016/0010-0285(83)90002-6

Haskell, R. E. (2001). *Transfer of learning: Cognition, instruction, and reasoning*. San Diego, CA: Academic Press.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory and Cognition*, *15*, 332–340. doi:10.3758/BF03197035

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, *27*, 317–326. doi:10.1007/s10648-015-9309-3

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*, 772–775. doi:10.1126/science.1199327

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968. doi:10.1126/science.1152408

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education*, *43*, 21–27. doi:10.1111/j.1365-2923.2008.03245.x

Lee, H. S., Betts, S., & Anderson, J. R. (2017). Embellishing problem-solving examples with deep structure information facilitates transfer. *Journal of Experimental Education*, *85*, 309–333. doi:10.1080/00220973.2016.1180277

Lim, S. W. H., Ng, G. J. P., & Wong, G. Q. H. (2015). Learning psychological research and statistical concepts using retrieval-based practice. *Frontiers in Psychology: Educational Psychology*, *6*, 1484.

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science*, *20*, 516–522. https://doi.org/10.1111/j.1467-9280.2009.02325.x

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, *27*, 360–372. doi:10.1002/acp.2914

McKeough, A., Lupart, J., & Marini, A. (Eds.). (1995). *Teaching for transfer: Fostering generalization in learning*. Mahwah, NJ: Lawrence Erlbaum.

Reed, S. K., Dempster, A., & Ettinger, M. (1985). Usefulness of analogous solutions for solving algebra word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 106–125.

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*, 382–395.

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

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

Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition*, *1*, 242–248. doi:10.1016/j.jarmac.2012.09.002

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. *Educational Psychologist*, *24*, 113–142. doi:10.1207/s15326985ep2402_1

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education*, *36*, 404–411.

Tempel, T., & Kubik, V. (2017). Test-potentiated learning of motor sequences. *Memory*, *25*, 326–334. doi:10.1080/09658211.2016.1171880

Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*, 247–264. doi:10.1007/s10648-015-9310-x

Yong, P. Z., & Lim, S. W. H. (2016). Observing the testing effect using Coursera video-recorded lectures: A preliminary study. *Frontiers in Psychology: Cognition*, *6*, p. 2064. doi:10.3389/fpsyg.2015.02064

Appendix A

Passage “sprint” administered during Phase 1:

A hypothesis is a prediction intended to be tested in a research study. Suppose the following hypothesis-testing scenario:

Does drinking coffee enable sportsmen to sprint faster?

There are three main steps to take in hypothesis testing:

Step 1

We restate the question as a research hypothesis and a null hypothesis about the populations.

The populations are defined as follows:

Population 1: *Sportsmen who drank coffee.*

Population 2: *Sportsmen who didn’t drink coffee.*

The goal is to discover whether the sprinting performance of sportsmen who consumed coffee is superior to the performance of sportsmen who did not consume coffee.

The research and null hypotheses, written in symbols, are as follows:

Research hypothesis: $\mu_1 < \mu_2$

Null hypothesis: $\mu_1 = \mu_2$

μ stands for the population mean (average). In this case, μ_1 stands for the average sprinting time (in seconds) of sportsmen who drank coffee, while μ_2 stands for the average sprinting time (in seconds) of those who didn’t drink coffee.

If our research hypothesis were true, then $\mu_1 < \mu_2$. In other words, sportsmen who drank coffee would sprint faster.

If coffee doesn’t affect sprinting performance, then the null hypothesis stands. In other words, sportsmen’s sprinting performance remains the same regardless of whether they drank coffee.

Take particular note of one possibility: Sportsmen who drank coffee might actually end up performing *worse* (i.e., yielding longer sprinting time) than those who did not drink coffee (yielding shorter sprinting time)—that is, $\mu_1 > \mu_2$. Such an outcome does not support our research hypothesis. Instead, such an outcome subsumes under our null hypothesis.

The critical idea is that in hypothesis testing, the research hypothesis and null hypothesis, when taken together, must encompass all possible outcomes. Specifically in this instance, the null hypothesis is true when (a) coffee did not affect sprinting performance or (b) coffee actually hampered sprinting performance. Thus, our research and null hypotheses should look like these, respectively:

Research hypothesis: $\mu_1 < \mu_2$

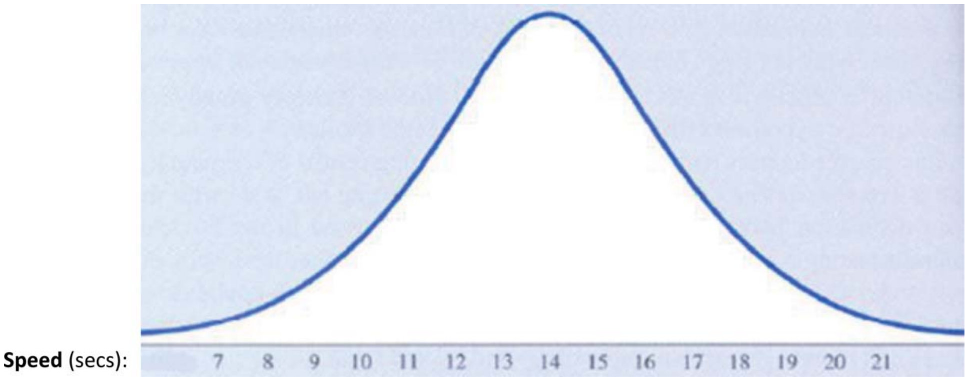
Null hypothesis: $\mu_1 \geq \mu_2$

Step 2

We have to determine the characteristics of the comparison population.

Specifically, before we are able to verify whether drinking coffee enhanced sportsmen's sprinting performance (in Population 1), we have to first find out the baseline performance of those (in Population 2) who did not drink coffee.

Let's take the following graph to depict Population 2's overall sprinting performance:



Most sportsmen in this noncoffee group took 14 seconds to complete their sprinting; most of their timings range between 11 and 17 seconds. Only very few took 8 seconds or less.

Here, we are interested to find out whether sportsmen who consumed coffee (in Population 1) would yield sprinting times that are very extreme (i.e., extremely faster), compared to those who did not consume coffee (in Population 2).

Step 3

We determine the extreme cutoff score on the comparison population, beyond which the research hypothesis should be accepted.

Suppose only the top 1% of sportsmen who did not drink coffee (in Population 2) took only 7 seconds or less. In other words, it is extremely rare for sportsmen, on an ordinary day, to demonstrate such outstanding sprinting performance.

Our goal is to determine whether drinking coffee is able to significantly enhance sprinting performance, such that it becomes this extremely good. Accordingly, we take 7 seconds as the extreme cutoff point, and consider this point as reflecting very extreme (extremely good) performance.

If sportsmen consumed coffee (in Population 1) and yielded outstanding sprinting performance that is below 7 seconds, the performance is taken as extremely good and we conclude that coffee significantly enhanced sprinting performance. In other words, we accept our research hypothesis (coffee enhances sprinting performance) as true.

However, if sportsmen consumed coffee (in Population 1) and their performance did not fulfil our cutoff score (i.e., they were not faster than 7 seconds), we conclude that coffee did not manage to significantly enhance sprinting performance. In other words, we cannot accept our research hypothesis as true and—- have to continue embracing the null hypothesis (coffee did not significantly enhance sprinting performance) as true.

Appendix B

Passage "sleep" administered during Phase 1:

A hypothesis is a prediction intended to be tested in a research study. Suppose the following hypothesis-testing scenario:

Does drinking milk before one sleeps enhance sleep quality?

There are three main steps to take in hypothesis testing:

Step 1

We restate the question as a research hypothesis and a null hypothesis about the populations.

The populations are defined as follows:

Population 1: *People who drank milk before sleeping.*

Population 2: *People who didn't drink milk before sleeping.*

The goal is to discover whether the sleep quality of those who consumed milk before sleeping is superior to the sleep quality of those who did not consume milk before sleeping.

The research and null hypotheses, written in symbols, are as follows:

Research hypothesis: $\mu_1 > \mu_2$

Null hypothesis: $\mu_1 = \mu_2$

μ stands for the population mean (average). In this case, μ_1 stands for the average sleep quality index scores (on a 0–20 scale, with 20 being the best) of people who drank milk before sleeping, while μ_2 stands for the average sleep quality index scores (on a 0–20 scale, with 20 being the best) of those who did not drink milk before sleeping.

If our research hypothesis were true, then $\mu_1 > \mu_2$. In other words, people who drank milk before sleeping had better sleep quality.

If milk does not affect sleep quality, in other words, one's sleep quality remains the same regardless of whether they drank milk before sleeping, then the null hypothesis stands. Take particular note of one possibility: People who drank milk before sleeping might actually end up with *poorer* sleep quality (i.e., yielding a poorer sleep quality index) than those who did not drink milk before sleeping (yielding a better sleep quality index)— that is, $\mu_1 < \mu_2$. Such an outcome will not support our research hypothesis. Instead, such an outcome subsumes under our null hypothesis.

The critical idea is that in hypothesis testing, the research hypothesis and null hypothesis, when taken together, must encompass all possible outcomes. Specifically in this instance, the null hypothesis is true when (a) milk did not affect sleep quality or (b) milk actually decreased sleep quality. Thus, our research and null hypotheses should look like these, respectively:

Research hypothesis: $\mu_1 > \mu_2$

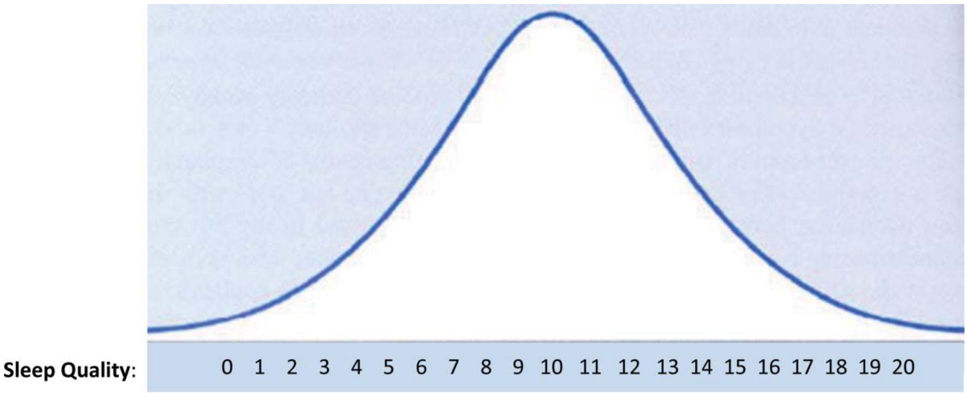
Null hypothesis: $\mu_1 \leq \mu_2$

Step 2

We have to determine the characteristics of the comparison population.

Specifically, before we are able to verify whether drinking milk enhanced one's sleep quality (in Population 1), we have to first find out the baseline performance of those (in Population 2) who did not drink milk.

Let's take the following graph to depict Population 2's overall sleep quality:



Most people in this non-milk-drinking group had a sleep quality index of 10; for most, sleep quality indexes were between 6 and 14. Very few had a sleep quality index of 17 or more.

Here, we are interested to find out whether people who consumed milk before sleeping (in Population 1) would yield sleep quality indexes that are very extreme (i.e., extremely higher) compared to those who did not consume milk before sleeping (in Population 2).

Step 3

We determine the extreme cutoff score on the comparison population, beyond which the research hypothesis should be accepted.

Suppose only the top 1% of people who did not drink milk before sleeping (in Population 2) had a sleep quality index of 19 or more. In other words, it is extremely rare for someone, on an ordinary day, to demonstrate such outstanding sleep quality. Our goal is to determine whether drinking milk before sleeping is able to significantly enhance sleep quality, such that it becomes this extremely good. Accordingly, we take a sleep quality index of 19 as the extreme cutoff point and consider this as reflecting very extreme (extremely good) sleep quality.

If people consumed milk before sleeping (in Population 1) and yielded an outstanding sleep quality index greater than 19, the sleep quality is taken as extremely good and we conclude that consuming milk before sleeping significantly enhanced sleep quality. In other words, we accept our research hypothesis (milk enhances sleep quality) as true.

However, if people consumed milk before sleeping (in Population 1) and their performance did not fulfil our cutoff score (i.e., they had a sleep quality of less than 19), we conclude that milk did not manage to significantly enhance sleep quality. In other words, we cannot accept our research hypothesis as true, but have to continue to embrace the null hypothesis (drinking milk before sleeping did not significantly enhance sleep quality) as true.

Appendix C

Final test question:

Based on what you have read and learned just now (or last week), how would you test the following hypothesis?

Does drinking coffee enable sportsmen to sprint faster?

OR

Based on what you have read and learned just now (or last week), how would you test the following hypothesis?

Does drinking milk before one sleeps enhance sleep quality?