


Cognitive Psychology

Repeated Retrieval Practice to Foster Students' Critical Thinking Skills

Lara M. van Peppen¹ ^a, Peter P. J. L. Verkoeijen², Anita Heijltjes³, Eva Janssen⁴, Tamara van Gog⁴

¹ Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Rotterdam, the Netherlands; Institute of Medical Education Research, Erasmus University Medical Center Rotterdam, Rotterdam, the Netherlands, ² Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Rotterdam, the Netherlands; Learning and Innovation Center, Avans University of Applied Sciences, Breda, the Netherlands, ³ Learning and Innovation Center, Avans University of Applied Sciences, Breda, the Netherlands, ⁴ Department of Education, Utrecht University, Heidelberglaan, the Netherlands

Keywords: repeated retrieval practice, spacing effect, retrieval practice, unbiased reasoning, critical thinking

<https://doi.org/10.1525/collabra.28881>

Collabra: Psychology

Vol. 7, Issue 1, 2021

There is a need for effective methods to teach critical thinking. Many studies on other skills have demonstrated beneficial effects of practice that repeatedly induces retrieval processes (repeated retrieval practice). The present experiment investigated whether repeated retrieval practice is effective for fostering critical thinking skills, focusing on avoiding biased reasoning. Seventy-five students first took a pre-test. Subsequently, they were instructed on critical thinking and avoiding belief-bias in syllogistic reasoning and engaged in retrieval practice with syllogisms. Afterwards, depending on the assigned condition, they (1) did not engage in extra retrieval practice; (2) engaged in retrieval practiced a second time (week later); or (3) engaged in retrieval practiced a second (week later) and a third time (two weeks later). Two/three days after the last practice session, all participants took a post-test consisting of practiced tasks (to measure learning relative to the pre-test) and non-practiced (transfer) tasks. Results revealed no significant difference between the pretest and the posttest learning performance as judged by the mean total performance (MC-answers + justification), although participants were, on average, faster on the post-test than on the pre-test. Exploring performance on MC-answers-only suggested that participants did benefit from instruction/practice but may have been unable to justify their answers. Unfortunately, we were unable to test effects on transfer due to a floor effect, which highlights the difficulty of establishing transfer of critical thinking skills. To the best of our knowledge, this is the first study that addresses repeated retrieval practice effects in the critical thinking domain. Further research should focus on determining the preconditions of repeated retrieval practice effects for this type of tasks.

1. Introduction

One of the most valued and sought after skills that higher education students are expected to learn is critical thinking (CT). CT is key to effective thinking about difficult issues, weighing evidence, determining credibility, and acting rationally, which is essential for succeeding in future careers and to be efficacious citizens (Billings & Roberts, 2014; Davies, 2013; Halpern, 2014; Van Gelder, 2005). The concept of CT can be expressed in a variety of definitions, but at its core, CT is “good thinking that is well reasoned and well supported with evidence” (H. A. Butler & Halpern, 2020, p. 152). One key aspect of CT is the ability to avoid biases in reasoning and decision-making (e.g., West et al., 2008), referred to as *unbiased reasoning*. Bias is said to occur when people rely on heuristics (i.e., mental shortcuts) dur-

ing reasoning prior to choosing actions and estimating probabilities that result in systematic deviations from ideal normative standards (i.e., derived from logic and probability theory: Stanovich et al., 2016; Tversky & Kahneman, 1974). As biased reasoning can have serious consequences in both daily life and complex professional environments, it is essential to teach CT in higher education (e.g., Koehler et al., 2002).

Not surprisingly, therefore, there is a growing body of literature on how to teach CT, including unbiased reasoning (e.g., Abrami et al., 2014; Heijltjes et al., 2015; Heijltjes, Van Gog, & Paas, 2014; Heijltjes, Van Gog, Leppink, et al., 2014; Janssen, Mainhard, et al., 2019; Janssen, Meulendijks, et al., 2019; Kuhn, 2005; Sternberg, 2001; Van Brussel et al., 2020; Van Peppen et al., 2018; Van Peppen, Verkoeijen, Heijltjes, et al., 2021; Van Peppen, Verkoeijen, Kolenbran-

a l.vanpeppen@erasmusmc.nl

der, et al., 2021). It is well established, for instance, that explicit teaching of CT combined with practice on domain-relevant problems improves learning of general CT-skills (Abrami et al., 2008, 2014) and CT-skills required for unbiased reasoning specifically (Heijltjes et al., 2015; Heijltjes, Van Gog, & Paas, 2014; Heijltjes, Van Gog, Leppink, et al., 2014). Especially when students are exposed to authentic and sense-making problems (i.e., authentic instruction) or discuss specific problems cooperatively (i.e., dialogue) and when these instructional approaches are combined with one-on-one coaching/mentoring on students' CT (Abrami et al., 2014).

Nonetheless, while some effective instructional approaches for learning CT have been identified, it is still unclear which methods are most effective in supporting the ability to transfer what has been learned (Halpern & Butler, 2019; Heijltjes et al., 2015; Heijltjes, Van Gog, & Paas, 2014; Heijltjes, Van Gog, Leppink, et al., 2014; Ritchhart & Perkins, 2005; Tiruneh et al., 2014, 2016; Van Peppen et al., 2018; Van Peppen, Verkoeijen, Heijltjes, et al., 2021; Van Peppen, Verkoeijen, Kolenbrander, et al., 2021). Transfer is the process of applying one's prior knowledge or skills to related materials or some new context (e.g., Barnett & Ceci, 2002; Cormier & Hagman, 2014; Haskell, 2001; Perkins & Salomon, 1992; Salomon & Perkins, 1989). There are some insights into fostering transfer of CT-skills to isomorphic tasks (in this study referred to as learning; e.g., Heijltjes, Van Gog, Leppink, et al., 2014), but not into transfer to novel tasks that share underlying principles but have not been previously encountered (e.g., Heijltjes et al., 2015; Heijltjes, Van Gog, Leppink, et al., 2014; Van Peppen et al., 2018; Van Peppen, Verkoeijen, Heijltjes, et al., 2021; Van Peppen, Verkoeijen, Kolenbrander, et al., 2021). As it is crucial that students can successfully apply the CT-skills acquired at a later time and to novel contexts/problems and it would be unfeasible to train students on each and every type of reasoning bias they will ever encounter, more knowledge is needed into the conditions that not only yield learning of CT-skills but also transfer.

Previous research has demonstrated that to establish learning *and* transfer, learners have to *actively* construct meaningful knowledge from to-be-learned information, by mentally organizing it in coherent knowledge structures and integrating these with one's prior knowledge (Bassok & Holyoak, 1989; Fiorella & Mayer, 2016; Gick & Holyoak, 1983; Holland et al., 1989; Wittrock, 2010). This, in turn, can aid future problem solving (Kalyuga, 2011; Renkl, 2014; Van Gog et al., 2019): if a situation presents similar requirements and the learner recognizes them, they may select and apply the same or a somewhat adapted learned procedure to solve the problem. One of the strongest learning techniques known to promote the construction of meaningful knowledge structures, is having students retrieve to-be-learned material from memory, known as practice testing or retrieval practice (e.g., Dunlosky et al., 2013; Fiorella & Mayer, 2015, 2016; Roediger & Butler, 2011). The effect of retrieval practice seems to be extremely robust (for reviews, see Carpenter, 2012; Delaney et al., 2010; Moreira et al., 2019; Pan & Rickard, 2018; Rickard & Pan, 2017; Roediger & Butler, 2011; Roediger & Karpicke, 2006b; Rowland, 2014) emerging on measures of both learning and transfer,

and with different kinds of materials and test formats (e.g., A. C. Butler, 2010; Carpenter & Kelly, 2012; McDaniel et al., 2012, 2013; Rohrer et al., 2010).

1.1 Repeated Retrieval Practice

The effect of retrieval practice seems to be positively related to the number of successful retrieval attempts during practice (e.g., Rawson & Dunlosky, 2011; Roediger & Karpicke, 2006a), albeit with diminishing returns. For example, in Experiment 2 from the study by Roediger and Karpicke (2006a), participants either studied a prose passage multiple times (SSSS condition), studied a prose passage multiple times and took one free recall retrieval practice test (SSST condition) or studied a prose passage once and took a free recall retrieval practice test thrice (STTT condition). Subsequently, a delayed final free recall test on the prose passage was administered in all conditions. The results on this final free recall test showed that taking a single retrieval practice test increased the free recall performance relative to the control condition from a mean score of 40% correct to a mean score of 56% correct. Furthermore, repeated retrieval practice (i.e., the STTT condition) increased the free recall performance to a mean of 61% correct, hence showing diminishing returns for extra retrieval practice. That is, where a single retrieval practice test in the SSST condition lifted final test performance with 16% points, the two additional retrieval practice tests increased the final test performance with only 5% points. These diminishing returns of repeated retrieval practice might be due to the fact that the practice testing effect depends not only on the number of successful retrieval attempts but also on the effort that is required to successfully retrieve information from memory. According to the retrieval effort hypothesis (e.g., Pyc & Rawson, 2009) the effect of retrieval practice becomes larger when successful retrieval attempts require more effort. When information is repeatedly retrieved from memory, the effort associated with successful retrieval is likely to decrease, which will lead to diminishing returns of repeated retrieval practice.

Despite the potential of repeated retrieval for learning, its impact has not been investigated in research on CT. Therefore, the present study sought to determine whether repeated retrieval practice is beneficial to foster learning of CT-skills as well, and whether it can additionally facilitate transfer. For educational practice, it is relevant to identify the most efficient schedule from among those that achieve a desired level of durability. While the majority of studies were conducted in laboratory settings, the current study was conducted as part of an existing CT-course using educationally relevant practice sessions (multiple practice tasks within a session) and retention intervals (days/weeks). To the best of our knowledge, this is the first study that investigated the effects of repeated retrieval practice in the CT-domain.

1.2 The Present Study

Participants first completed a pretest including syllogistic reasoning tasks (for an overview of the study design, see [figure 1](#)), which examined their tendency to be influenced by the believability of a conclusion when evaluating the

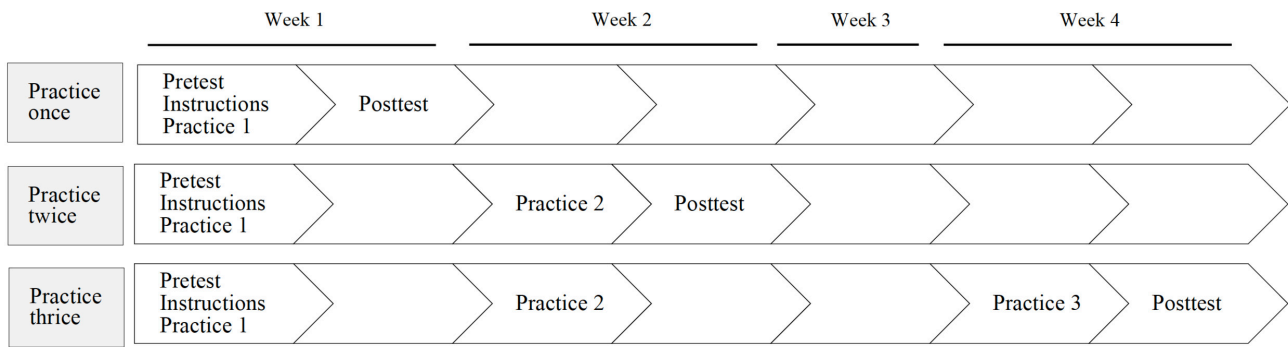


Figure 1. Study design per instructional condition

Note. All participants completed the posttest two or three days after their last practice session.

logical validity of arguments. Thereafter, they received instructions on CT in general and on syllogisms in particular. Subsequently, they engaged in retrieval practice with these tasks on domain-specific problems. Depending on condition, participants (1) did not engage in extra retrieval practice with these tasks (practice once); (2) engaged in retrieval practice a second time (one week later; practice twice); or (3) engaged in retrieval practice a second (one week later) and third time (two weeks after second time; practice thrice). Subsequently, all participants completed a posttest including practiced tasks (i.e., syllogistic reasoning tasks; measure of *learning*) and non-practiced tasks (i.e., Wason selection tasks; measure of *transfer*) two or three days after their last practice session. Participants had to indicate after each test and practice item how much effort they invested on that item and time-on-task was logged during all phases. Furthermore, they were asked after each practice session to assess how well they thought they understood the practice problems (i.e., global judgment of learning; JOL) to gain insight into the added value of extra practice according to the students themselves. Previous research has demonstrated that students' JOLs are related to their learning strategies and study time (i.e., monitoring learning processes; e.g., Koriati, 1997; Nelson et al., 1994; Zimmerman, 2000) and, thus, may indirectly contribute to performance enhancement.

We hypothesized that explicit CT-instructions combined with retrieval practice would be effective for *learning*: thus, we expected an overall mean pretest to posttest performance gain on learning items in all conditions (Hypothesis 1). Furthermore, and more importantly, we expected that practicing retrieval twice would lead to a higher pretest to posttest performance gain on learning items (Hypothesis 2a) and a higher posttest performance on transfer items¹ (Hypothesis 3a) than practicing retrieval once. We expected that practicing retrieval thrice would lead to a higher pretest to posttest performance gain on learning items (Hypothesis 2b) and a higher posttest performance on transfer

items (Hypothesis 3b) than practicing retrieval twice. However, as outlined before, prior research suggests that additional retrieval practice will have diminishing returns on the final test, so we expected these differences to be smaller than the differences between practicing retrieval once and twice.

To get more insight into the effectiveness (higher performance) and efficiency (i.e., performance/investment of mental effort or time; Van Gog & Paas, 2008) of repeated retrieval practice on learning and transfer, we explored the invested mental effort, time-on-test, and JOLs. Thus, we exploratively compared the practice conditions on invested mental effort on test items, time-on-test, and JOLs.

2. Method

The hypotheses and complete method section were pre-registered on the Open Science Framework (OSF). All data, script files, and materials (in Dutch) are available on the project page that we created for this study (<https://osf.io/pfmyg/>).

2.1 Participants and Design

Participants were all first-year 'Safety and Security Management' students attending a Dutch University of Applied Sciences ($N = 103$). Eleven students did not complete the posttest and two students completed the posttest a week late and therefore were excluded from the analyses (as this may have influenced the results). Seventeen participants were excluded because of non-compliance, i.e., when more than half of the practice tasks during one of the essential practice sessions were not read seriously.² Due to a technical problem, one class of students (i.e., 24 students) did not receive the demographic questionnaire and the pretest. Together, this resulted in a final sample of 75 students for the posttest-only analyses (i.e., completed all essential sessions, excluding the demographic questions and pretest) and a subsample of 51 students (68%) for the pretest to

¹ Because transfer items were not included in the pretest, we are not able to detect transfer *gains*.

² Fast readers (i.e., maximum reading speed of 0.17 seconds per word; e.g., Trauzettel-Klosinski & Dietz, 2012), taken as a limit.

posttest analyses (i.e., completed all essential sessions: $M_{\text{age}} = 19.47$, $SD = 1.64$; 25 female).

We calculated power functions of our analyses using the G*Power software (Faul et al., 2009). The power of our one-way ANOVAs –under a fixed alpha level of .05 and with a sample size of 75– is estimated at .11, .47, and .87 for picking up a small ($\eta_p^2 = .01$), medium ($\eta_p^2 = .06$), and large ($\eta_p^2 = .14$) effect. Regarding the crucial interaction between number of practice sessions and test moment –again calculated under a fixed alpha level of .05, but with a sample size of 51 and a correlation between measures of .64– the power is estimated at .27, .95, and >.99 for picking up a small, medium, and large interaction effect, respectively. Thus, our sample size under the above assumptions should be sufficient to pick up medium-large effects, and previous studies on repeated (retrieval) practice mainly demonstrated medium-large effects (e.g., Roediger & Karpicke, 2006b).

The educational committee of the university approved on conducting this study within the curriculum. In week 1, all participants first completed the CT-skills pretest, followed by the CT-instructions and practice session one (see [figure 1](#) for an overview). Participants were randomly assigned to one of three conditions. They either (1) did not practice extra with the tasks (practice once condition, posttest only: $n = 26$; both tests: $n = 16$), (2) practiced a second time in week 2 (practice twice condition, $n = 25$; $n = 16$), or (3) practiced a second time in week 2 and a third time in week 4 (practice thrice condition, $n = 24$; $n = 19$). Participants completed the CT-skills posttest two or three days after their last practice session.

2.2 Materials

2.2.1 CT-skills tests. The content of the surface features of all items was adapted to participants' study domain. The pretest consisted of 16 syllogistic reasoning items across two categories (i.e., conditional and categorical syllogisms, see Appendix S1 for an example with explanation of each category), which were used to measure *learning*, as these were instructed and practiced during the training phase. All of the items included a belief bias (i.e., when the conclusion aligns with your prior beliefs or real-world knowledge but is invalid or vice versa; Evans et al., 1983; Markovits & Nantel, 1989; Newstead et al., 1992) and examined the tendency to be influenced by the believability of a conclusion when evaluating the logical validity of arguments (Evans, 1977, 2003). These types of tasks are frequently used to measure people's ability to avoid biases (e.g., Stanovich et al., 2016).

Our tests consisted of 3 × affirming the consequent of a conditional statement (if p then q , q therefore p ; invalid); 3 × denying the consequent of a conditional statement (if p then q , not q therefore not p ; valid); 2 × affirming the antecedent of a conditional statement (if p then q , p therefore q ; valid); 2 × denying the antecedent of a conditional statement (if p then q , not p therefore not q ; invalid); 3 × categorical syllogism 'no A is B, some C are B, therefore some C are not A' (valid); and 3 × categorical syllogism 'no A is B, some C are B, therefore some A are not C' (invalid). Participants had to indicate for each item whether the conclusion was valid or invalid and to explain their multiple-choice (MC)

answer to check their understanding (on the MC-answers they might be guessing). They could earn 1 point for the correct MC-answer and 1 point for a correct and 0.5 point for a partially correct explanation (see subsection 2.4). The MC and explanation scores were sum-scored and, thus, the maximum total score on the learning items was 32 points.

The posttest was identical to the pretest but, additionally, six Wason selection items were added that measured the tendency to confirm a hypothesis rather than to falsify it (see the Appendix for two examples with explanations; e.g., Dawson et al., 2002; Evans, 2002; Stanovich, 2011). These items measured *transfer* as they were not instructed/practiced but shared similar features with the four types of conditional syllogisms. Our test consisted of 3 abstract versions and 3 versions including study-related context. A MC-format with four answer options was used in which only a specific combination of two selected answers was the correct answer. One point was assigned for each correct answer (see subsection 2.4), resulting in a maximum total score of six points on the transfer items.

2.2.2 CT-instructions. The video-based CT-instructions (15 min.) consisted of a general CT-instruction (i.e., features of CT and attitudes/skills needed to think critically) and explicit instructions on belief-bias in syllogisms that consisted of a worked example of each of the six types in the pretest. The worked examples showed the correct line of reasoning and included possible problem-solving strategies, which allowed participants to mentally correct initially erroneous responses. At the end, participants received a hint stating that the principles used in these examples can be applied with several other reasoning tasks.

2.2.3 CT-practice. Participants could practice retrieval on the six types of syllogisms on topics that they might encounter in their working-life. Participants were instructed to read the problems thoroughly and to choose the correct MC-answer option, provided directly below the problems. They had to deliberately recall the relevant information from their memory to solve the problems. After each practice-task, they received correct-answer feedback and were given a worked example in which the line of reasoning was explained in steps and clarified with a visual representation. The second and third practice sessions were parallel versions of the first one (i.e. structurally equivalent problems but with different surface features).

2.2.4 Mental effort. After each test item and after each CT-practice problem, participants were asked to indicate how much effort they invested on completing that task, on a 9-point scale ranging from (1) very, very low effort to (9) very, very high effort (Paas, 1992).

2.2.5 Global judgments of learning (JOL). At the end of each practice session, participants made a JOL on how well they thought they understood the CT-practice problems on a 7-point scale ranging from (1) very poorly to (7) very well (Koriat et al., 2002; Thiede et al., 2003).

2.3 Procedure

The study was run during the first four weeks of a CT-course in the Integral Safety and Security Management study program of an institute of higher professional education. The CT-skills pretest and first practice session were

conducted during the first lesson in a computer classroom at the participants' university with an entire class of students and their teacher present. The extra practice sessions and the posttest were completed entirely online (cf. Heijltjes, Van Gog, & Paas, 2014). Participants came from four different classes and within each class, students were randomly assigned to one of the conditions. All materials were delivered in a computer-based environment (Qualtrics platform). Participants could work at their own pace, were allowed to use scrap paper while solving the tasks, and time-on-task was logged during all phases.

In advance of the first lesson, the students were informed by their teacher about the experiment (i.e., procedure and time window). When entering the classroom in *week 1*, participants were instructed to sit down at one of the desks and read the A4-paper containing some general instructions and a link to the Qualtrics environment where they first had to sign an informed consent form. Thereafter, they had to fill in a demographic questionnaire and complete the pretest. After each test item, they had to indicate how much mental effort they invested. Subsequently, participants entered the practice phase in which they first viewed the video-based CT-instructions (15 min), followed by the practice tasks. At the end of the practice phase, participants had to indicate their JOL. Participants had to wait (in silence) until the last participant had finished before they were allowed to leave the classroom.

One day before each online session (i.e., practice session 2 and 3 and posttest), participants received an e-mail with a reminder and the request to reserve time for this mandatory part of their CT-course. One hour before participants could start, they received the link to the Qualtrics environment. They were given a specific time window (8 am to 10 pm that day) to complete these sessions. Two or three days after session 1, participants of the practice once condition had to complete the posttest. In the beginning of *week 2*, all participants had to complete the second practice session. Since the content of our materials was part of the final exam of this course and the ethical guidelines of the institute of higher professional education state that all students should have been offered the same exam materials, participants of the practice once condition practiced with the extra practice materials but they were no longer included in the experiment. Two or three days after session 2, participants of the practice twice condition had to complete the posttest. Due to practical reasons (i.e., one week school holiday), the procedure of week 2 was repeated in *week 4*; all participants had to complete the third practice session but students in the practice once and twice conditions were no longer partaking in the experiment and those in the practice thrice condition had to complete the posttest after three days. Participants who did not complete either the posttest or one of the extra practice sessions received an e-mail the day after the specific time-window with the message that they could complete it that day as a last opportunity.

2.4 Data Analysis

Items were scored for accuracy; 1 point for each correct MC-alternative and a maximum of 1 point (increasing in steps of 0.5) for the correct explanation on the learning

items (coding scheme can be found on our OSF-page). Unfortunately, one transfer item had to be removed from the test due to incorrectly offered MC-answer options. As a result, participants could attain a maximum total score of 32 points on the learning items and five points on the transfer items. For comparability, learning and transfer outcomes were computed as percentage correct scores instead of total scores. Two raters independently scored 25% of the explanations on the learning items of the posttest. Intraclass correlation coefficient (two-way mixed, consistency, single-measures; McGraw & Wong, 1996) was 0.996, indicating excellent interrater reliability (Koo & Li, 2016). The remainder of the tests was scored by one rater. Cronbach's alpha was .74 on the learning items on the pretest, .71 on the learning items on the posttest and .79 on the transfer items.

Boxplots were created to identify outliers (i.e., values that fall more than 1.5 times the interquartile range above the third quartile or below the first quartile) in the data. If any, we first conducted the analyses on the data of all participants and reran the analyses on the data without outliers. If outliers had influence on the results, we reported the data of both analyses. If not, we only reported the results on the full data set. In case of severe violations of the assumption of normality for our analyses, we conducted appropriate non-parametric tests.

3. Results

For all analyses in this paper, a p -value of .05 was used as a threshold for statistical significance. Partial eta-squared (η^2) is reported as an effect size for all ANOVAs with $\eta^2 = .01$, $\eta^2 = .06$, and $\eta^2 = .14$ denoting small, medium, and large effects, respectively (Cohen, 1988). Cramer's V is reported as an effect size for chi-square tests with (having 2 degrees of freedom) $V = .07$, $V = .21$, and $V = .35$ denoting small, medium, and large effects, respectively.

3.1 Check on Condition Equivalence

Before running any of the main analyses, we checked our conditions on equivalence. Preliminary analyses confirmed that there were no a-priori differences between the conditions in age, $F(2, 50) = 0.46$, $p = .634$, $\eta^2 = .02$; educational background, $\chi^2(8) = 12.69$, $p = .12$, $V = .35$; performance on the pretest, $F(2, 47) = 0.24$, $p = .790$, $\eta^2 = .01$; time spent on the pretest, $F(2, 47) = 0.74$, $p = .481$, $\eta^2 = .03$; mental effort invested on the pretest, $F(2, 47) = 0.82$, $p = .445$, $\eta^2 = .03$; performance on practice problems session one, $F(2, 74) = 0.12$, $p = .889$, $\eta^2 < .01$; time spent on practice problems session one, $F(2, 74) = 0.89$, $p = .417$, $\eta^2 = .02$; effort invested on practice problems session one, $F(2, 74) = 0.47$, $p = .629$, $\eta^2 = .01$; and global JOL, $F(2, 74) = 0.36$, $p = .701$, $\eta^2 = .01$. We found a gender difference between the conditions, $\chi^2(2) = 6.23$, $p = .043$, $V = .35$. However, gender did not correlate significantly with any of our performance measures (minimum $p = .669$) and was therefore not a confounding variable.

Table 1. Means (SD) of Test performance on learning items (% correct score), Test performance on transfer items (% correct score), Invested mental effort during test (1–9), Time-on-task during test (in seconds), and Global Judgment of Learning (1–7) after the last practice session per Instructional condition

| | | | Instructional conditions | | | |
|-----------------------------|----------|----|--------------------------|---------------|----------------|-----------------|
| | | | N | Practice once | Practice twice | Practice thrice |
| Test performance | | | | | | |
| Learning items | Pretest | 51 | 43.85 (18.22) | 47.36 (21.07) | 45.23 (17.25) | |
| | Posttest | 51 | 47.17 (14.46) | 51.37 (18.94) | 49.01 (14.94) | |
| | Posttest | 75 | 47.06 (15.88) | 48.56 (17.85) | 47.92 (14.22) | |
| Transfer items | Posttest | 75 | 7.69 (19.66) | 5.60 (17.81) | 1.67 (8.16) | |
| Mental effort | | | | | | |
| Learning items | Pretest | 51 | 4.32 (1.13) | 3.65 (1.28) | 3.82 (1.27) | |
| | Posttest | 51 | 4.48 (1.25) | 4.22 (1.47) | 4.22 (1.47) | |
| | Posttest | 75 | 4.64 (1.19) | 4.60 (1.19) | 4.39 (1.38) | |
| Transfer items | Posttest | 75 | 4.23 (1.43) | 4.35 (1.48) | 4.01 (1.68) | |
| Time on task | | | | | | |
| Learning items ^a | Pretest | 46 | 74.48 (17.13) | 75.63 (21.25) | 74.43 (18.57) | |
| | Posttest | 46 | 58.82 (31.86) | 51.35 (22.85) | 45.80 (23.36) | |
| | Posttest | 70 | 59.64 (21.89) | 58.82 (26.08) | 51.94 (28.50) | |
| Transfer items | Posttest | 75 | 38.04 (19.11) | 37.70 (25.47) | 29.04 (16.75) | |
| Global JOL | | 75 | 4.04 (1.64) | 4.76 (1.17) | 4.63 (1.47) | |

^a Means (SD) of the data excluding outliers.

3.2 Planned Analyses

We conducted pretest to posttest analyses on the data of participants who completed all essential experimental sessions ($n = 51$) and posttest-only analyses on the data of participants who missed the demographic questions and pretest ($n = 75$). Because of a floor effect on transfer performance, analysis of the transfer data would unfortunately not be very meaningful, and we therefore report only descriptive statistics on those data. Together with the descriptive statistics of the other dependent variables, these can be found in [Table 1](#).

3.2.1 Performance on learning items. In contrast to Hypotheses 1 and 2a, a 2×3 mixed ANOVA with Test Moment (pretest, posttest) as within-subjects factor and Condition (practice once, practice twice, practice thrice) as between-subjects factor on performance on learning items revealed no main effects of Test Moment, $F(1, 48) = 3.05$, $p = .087$, $\eta^2 = .06$, and Condition, $F(2, 48) = 0.24$, $p = .788$, $\eta^2 = .01$. Furthermore, there was no interaction between Test Moment and Condition, $F(2, 48) = 0.01$, $p = .991$, $\eta^2 < .01$. A one-way ANOVA with the full sample on the posttest data only, did not reveal an effect of Condition either, $F(2, 72) = 0.06$, $p = .945$, $\eta^2 < .01$.

3.2.2 Mental effort. A 2×3 mixed ANOVA on invested mental effort on the learning items, with Test Moment (pretest, posttest) as within-subjects factor and Condition (practice once, practice twice, practice thrice) as between-subjects factor showed a main effect of Test Moment, $F(1,$

$48) = 8.41$, $p = .006$, $\eta^2 = .15$; less effort was invested on learning items on the pretest ($M = 3.93$, $SD = 1.24$) than the posttest ($M = 4.32$, $SD = 1.30$). There was no main effect of Condition, $F(2, 48) = 0.67$, $p = .515$, $\eta^2 = .03$, nor an interaction between Test Moment and Condition, $F(2, 48) = 0.85$, $p = .435$, $\eta^2 = .03$. A one-way ANOVA with the full sample on the posttest data only, did not reveal an effect of Condition either, $F(2, 72) = 0.28$, $p = .754$, $\eta^2 = .01$.

3.2.3 Time-on-test. Because the data was not normally distributed, we conducted a Kruskal-Wallis H test with Condition (practice once, practice twice, practice thrice) as between-subjects factor on pretest-posttest differences in time spent on learning items. The results showed that there was no significant difference between conditions in pretest-posttest time spent on learning items, $\chi^2(2) = 1.54$, $p = .464$, $\eta^2 = .01$. A Kruskal-Wallis H test on the posttest-only data with Condition (practice once, practice twice, practice thrice) as between-subjects factor, showed that there was no significant difference in time spent on posttest learning items between conditions, $\chi^2(2) = 4.54$, $p = .103$, $\eta^2 = .04$. In addition to the results of the analysis on the full data, a 2×3 mixed ANOVA on the data without five outliers with Test Moment (pretest, posttest) as within-subjects factor and Condition (practice once, practice twice, practice thrice) as between-subjects factor did reveal a significant effect of Test Moment, $F(1, 42) = 39.34$, $p < .001$, $\eta^2 = .48$; more time was spent on the pretest ($M = 73.84$, $SD = 17.55$) than the posttest ($M = 49.26$, $SD = 21.14$).

3.2.4 Global judgments of learning. Finally, we examined differences in global JOLs using a one-way ANOVA. The results revealed no main effect of Condition, $F(2, 74) = 1.82, p = .170, \eta^2 p^2 = .05$.

3.3 Exploratory Analyses

To gain more insight into the effects of repeated retrieval practice, we explored participants' level of performance during practice session one, two, and three.³ Descriptive statistics showed that on average, performance increased with increasing practice opportunities: mean percentage correct during practice session one was 58.67% ($SD = 21.29; n = 75$), during session two 65.31% ($SD = 19.20; n = 49$), and during practice three 69.44% ($SD = 16.79; n = 24$).⁴ Since the transfer items of the tests shared similar features with the four types of conditional syllogisms, we additionally explored participants' level of performance during learning on these types only. Again, descriptive statistics showed that performance increased: mean percentage correct during practice session one was 55.33% ($SD = 24.42; n = 75$), during practice session two 63.78% ($SD = 25.55; n = 49$), and during practice session three 69.79% ($SD = 19.48; n = 24$).

Additionally, we explored whether performance on MC-questions only on the syllogism (learning) items improved after instruction and practice, using a 2×3 mixed ANOVA with Test Moment (pretest, posttest) as within-subjects factor and Condition (practice once, practice twice, practice thrice) as between-subjects factor. The results indeed revealed a main effect of Test Moment, $F(1, 47) = 20.26, p < .001, \eta^2 p^2 = .30$; performance was better on the posttest ($M = 68.66, SE = 2.30$) than the pretest ($M = 57.42, SE = 2.60$). There was, however, no significant main effect of Condition, $F(2, 47) = 0.50, p = .613, \eta^2 p^2 = .02$, nor an interaction between Test Moment and Condition, $F(2, 47) = 0.01, p = .990, \eta^2 p^2 < .01$. Finally, we explored how much time participants spent on the worked-example feedback after correct and incorrect retrievals. Both test and descriptive statistics (see Table 2) showed that participants spent – with almost all practice tasks – more time on the worked-example feedback after incorrect retrievals than after correct retrievals. Although participants generally spent less time on the worked-example feedback as they practiced more often (i.e., during a later practice session), this pattern is found during each of the three practice sessions.

3.4 Addressing Potential Power Issues

Due to a technical problem, our final sample was considerably smaller than predetermined and might have been insufficient to detect a small-medium interaction effect. Since adding participants to an already completed experiment will increase the Type 1 rate (alpha) and conducting

a second identical experiment (i.e., in the context of an actual course) would be resource-demanding, we decided to exploratory apply whether or not that would be worthwhile, using a sequential stopping rule (SSR: see, for example Arghami & Billard, 1982, 1991; Botella et al., 2006; Doll, 1982; Fitts, 2010; Pocock, 1992; Ximénez & Revuelta, 2007). SSRs make it possible to stop early when statistical significance is unlikely to be achieved with the planned number of participants.

One SSR that is simple, efficient, and appropriate to this experiment is the COAST (composite open adaptive stopping rule; Frick, 1998). The COAST allows to stop testing participants and reject the null hypothesis if the p -value is less than a lower criterion of .01; to stop testing participants and retain the null hypothesis if the p -value is greater than an upper criterion of .36; and to test more participants if the p -value is between these two values. In the present study, the p -values of our main analyses (i.e., on performance measures) were obviously larger than the high criterion of .36. Hence, there was no hint of an existing effect of repeated retrieval practice in the present study and, thus, we decided not to add additional participants.

4. Discussion

The current study investigated whether repeated retrieval practice is beneficial to foster learning of CT-skills and whether it can additionally facilitate transfer. Contrary to our expectations, we did not reveal pretest to posttest performance gains on learning items. Thus, we did not replicate the finding that participants' performance improves after explicit instructions combined with retrieval practice on domain-specific problems (Hypothesis 1: e.g., Heijltjes et al., 2015; Van Peppen et al., 2018; Van Peppen, Verkoeijen, Heijltjes, et al., 2021; Van Peppen, Verkoeijen, Kolenbrander, et al., 2021). It should be noted, however, that this comparable level of posttest performance was attained in less time than pretest performance (i.e., prior to instruction/practice). Moreover, our exploratory findings on performance on MC-questions only, suggest that students did benefit from instructions and retrieval practice. This difference in outcomes when looking at MC-answers and total scores (i.e., MC + justification) could mean that participants did learn what the right answer was, but may have been unable to justify their answers sufficiently. In that case, however, our intervention only resulted in simple memorization (i.e., rote learning; Mayer, 2002) instead of a deeper understanding of the subject matter. This might perhaps also explain the occurrence of a floor effect on performance on transfer items, as transfer of knowledge or skills depends on how well-developed the knowledge structures are that are formed during initial learning (e.g., Perkins & Salomon, 1992).

³ This concerns all participants who engaged in the relevant practice sessions (i.e., all conditions in practice session one, practice twice and thrice in session two, and practice thrice in session three).

⁴ We additionally tested within the practice thrice condition ($n = 24$) whether there was a significant difference in performance during practice session one, two, and three. Performance increased on average with increasing practice opportunities ($M_1 = 60.42\%, M_2 = 65.97\%, M_3 = 69.44\%$), but these differences (possibly due to the small sample size) were not significant, $F(2, 46) = 1.94, p = .155, \eta^2 p^2 = .08$.

Table 2. Means (SD) of time spent on worked-example feedback after correct and incorrect retrievals (in seconds) and t-test statistics (df) per Practice task and Practice session.

| | Session one <i>n</i> = 75 | | | Session two <i>n</i> = 49 | | | Session three <i>n</i> = 24 | | |
|--------|------------------------------|---------------|-------------|------------------------------|---------------|-------------|--------------------------------|-------------------|------------|
| | Correct | Incorrect | t-test | Correct | Incorrect | t-test | Correct | Incorrect | t-test |
| Task 1 | 8.74 (8.74) | 24.14 (21.75) | 3.59 (73)* | 6.09 (6.01) | 18.47 (15.61) | 3.80 (47)** | 5.67 (5.55) | 13.05 (17.82) | 1.37 (22)* |
| Task 2 | 9.81 (9.16) | 23.71 (25.79) | 3.45 (73)* | 6.98 (14.10) | 30.41 (74.19) | 1.69 (47) | 4.21 (3.31) | 12.82 (11.70) | 2.90 (22)* |
| Task 3 | 6.15 (6.56) | 17.44 (11.25) | 4.56 (73)** | 12.34 (47.86) | 11.96 (14.23) | -0.02 (47) | 8.24 (11.48) | 0 ^a | |
| Task 4 | 9.05 (10.00) | 26.99 (23.95) | 3.31 (73)* | 10.84 (20.46) | 28.62 (63.37) | 1.23 (47) | 7.20 (10.37) | 22.89 (32.08) | 1.67 (22) |
| Task 5 | 9.45 (11.90) | 24.65 (22.55) | 3.38 (73)* | 7.21 (8.20) | 24.19 (48.24) | 2.23 (47)* | 5.39 (3.96) | 8.54 ^b | 0.78 (22) |
| Task 6 | 15.23 (16.23) | 32.24 (23.59) | 3.37 (72)* | 7.46 (10.65) | 22.72 (32.29) | 2.16 (47)* | 10.12 (13.38) | 19.96 (32.28) | 0.91 (22) |

^a None of the participants completed this task incorrectly. ^b Only one of the participants completed this task incorrectly.

p* < .05, *p* < .001.

In line with previous repeated retrieval findings (e.g., Roediger & Butler, 2011), average performance scores during practice seemed to increase with more repetitions. However, repeated retrieval practice did not have a significant effect –compared to practice once– on performance on the final test (i.e., on learning items; Hypotheses 2a/2b). Unfortunately, we were unable to test whether repeated retrieval practice would enhance transfer (Hypotheses 3a/3b) due to a floor effect. Because the power of our study was only sufficient to pick up medium-to-large effects of repeated retrieval, it could be that additional retrieval practice had an unidentifiable small effect. In the current study, each practice session consisted of multiple practice tasks (instead of one as in most studies) and it could, therefore, be argued that practice once in this study can already be seen as repeated practice, which possibly explains the absence of substantial effects of repeated retrieval.

Another potential explanation for the lack of effect of additional retrieval practice, might lie in the feedback that was provided after each retrieval attempt. While many studies only show a retrieval practice effect when feedback is provided (for an overview, see Van Gog & Sweller, 2015) and others show that elaborative feedback can enhance effects of retrieval practice (e.g., Pan et al., 2016; Pan & Rickard, 2018), findings from recent research suggest that the feedback after each retrieval attempt may have eliminated the repeated retrieval effect (Kliegl et al., 2019; Pastötter & Bäuml, 2016; Storm et al., 2014). According to the bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011), feedback only strengthens knowledge that is not successfully retrieved, whereas knowledge that is successfully retrieved is hardly affected by subsequent feedback. As such, it may be that participants in the condition that merely practiced once (i.e., lowest performance during practice) processed the feedback better and, therefore, performed equally well on the final test as participants in the other conditions. Moreover, it may be that participants' motivation to learn the correct answer was higher when they were unable to provide the correct answer during retrieval practice than when they were able to do so (e.g., Kang et al., 2009; Potts & Shanks, 2019). Our findings regarding time spent on worked-example feedback after correct/incorrect retrievals support this idea (i.e., more time spent after incorrect than correct retrievals). The possible elimination of a lag effect on learning problem-solving skills by providing feedback after each retrieval attempt is an interesting issue for future research.

Although participants achieved a considerably high level of performance during retrieval practice (approx. 60–70 percent correct), which was comparable to previous studies that did demonstrate beneficial effects of repeated retrieval practice (e.g., A. C. Butler, 2010; Roediger & Karpicke, 2006b), a floor effect on performance on transfer items had arisen. Since the practice tasks consisted of MC-questions only, this finding again supports the idea that students do benefit from instructions and retrieval practice but may have been unable to justify their answers on the tests sufficiently. Another likely cause for this floor effect may be that participants lacked profound in-depth understanding of the structural overlap between syllogisms and Wason selection tasks (i.e., measure of transfer). During practice, partici-

pants could earn one point for each correctly solved syllogism. Each transfer item, however, required recall and application of *all* four conditional syllogism principles to solve it correctly and, thus, to earn one point. Future studies on to-be-transferred problem-solving procedures as in the current study, should guarantee sufficient understanding of structural features of tasks and complete recall of the procedure during retrieval practice. It may be helpful to provide longer or more extensive practice, including more guidance in identifying how tasks are related. Potentially, practicing retrieval until all retrievals are successful and complete might be a solution for complete recall of procedures (i.e., successive relearning: e.g., Bahrck, 1979; Rawson et al., 2013). Given that transfer of CT skills from trained to untrained tasks remains elusive (as our current results also underline), there is an urgent need to determine the exact obstacles to the transfer of CT-skills, which could lie in a failure to recognize that the acquired knowledge is relevant to the new task, inadequate recall of the acquired knowledge, and/or difficulties in actually applying that knowledge onto the new task (i.e., three-step process of transfer; Barnett & Ceci, 2002).

To the best of our knowledge, this is the first study that investigated the effects of repeated retrieval practice in the CT-domain. Moreover, while the majority of research on repeated retrieval practice has been conducted in laboratory settings, the current was conducted as part of an existing CT-course –using educationally relevant practice sessions and retention intervals. As such, it adds to the small body of literature on what instructional designs are (or are not) efficient and effective for CT-courses aiming at learning and transfer of CT-skills, which is relevant for both educational science and educational practice.

Author Contributions

Contributed to conception and design: LVP, PV, AH, TVG
 Contributed to acquisition of data: LVP
 Contributed to analysis and interpretation of data: LVP, PV, TVG
 Drafted and/or revised the article: LVP, PV, AH, EJ, TVG
 Approved the submitted version for publication: LVP, PV, AH, EJ, TVG

Acknowledgments

The authors would like to thank Stefan V. Kolenbrander for his help with running this study and Esther Stoop and Marjolein Looijen for their assistance with coding the data.

Funding Information

This work was supported by The Netherlands Organisation for Scientific Research (project number 409-15-203). TVG, PV, and AH were involved in the funding acquisition. The funding source was not involved in this study/manuscript.

Competing Interests

The authors have no other competing interests to declare. PV is an editor at Collabra: Psychology. He was not involved in the review process of this article.

Data Accessibility Statement

All data, script files, and materials (in Dutch) are available on the project page that we created for this study (anonymized view-only link: <https://osf.io/pfmyg/>).

Supplemental Material

Appendix S1. Example Items CT-skills Tests. Docx
Peer Review History. Docx

Submitted: December 06, 2020 PDT, Accepted: September 20, 2021 PDT



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

REFERENCES

- Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2014). Strategies for Teaching Students to Think Critically: A Meta-Analysis. *Review of Educational Research*, 85(2), 275–314. <https://doi.org/10.3102/0034654314551063>
- Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research*, 78(4), 1102–1134. <https://doi.org/10.3102/0034654308326084>
- Arghami, N. R., & Billard, L. (1982). A modification of a truncated partial sequential procedure. *Biometrika*, 69(3), 613–618. <https://doi.org/10.1093/biomet/69.3.613>
- Arghami, N. R., & Billard, L. (1991). A partial sequential t-test. *Sequential Analysis*, 10(3), 181–197. <https://doi.org/10.1080/07474949108836234>
- Bahrnick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General*, 108, 296.
- Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin*, 128(4), 612–637. <https://doi.org/10.1037/0033-2909.128.4.612>
- Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(1), 153–166. <https://doi.org/10.1037/0278-7393.15.1.153>
- Billings, L., & Roberts, T. (2014). *Teaching critical thinking: Using seminars for 21st century literacy*. Routledge. <https://doi.org/10.4324/9781315854595>
- Botella, J., Ximénez, C., Revuelta, J., & Suero, M. (2006). Optimization of sample size in controlled experiments: The CLAST rule. *Behavior Research Methods, Instruments, & Computers*, 38(1), 65–76. <https://doi.org/10.3758/bf03192751>
- Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(5), 1118–1133. <https://doi.org/10.1037/a0019902>
- Butler, H. A., & Halpern, D. F. (2020). Critical Thinking Impacts Our Everyday Lives. In R. J. Sternberg & D. F. Halpern (Eds.), *Critical Thinking in Psychology* (pp. 152–172). Cambridge University Press. <https://doi.org/10.1017/9781108684354.008>
- Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, 21(5), 279–283. <https://doi.org/10.1177/0963721412452728>
- Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin & Review*, 19(3), 443–448. <https://doi.org/10.3758/s13423-012-0221-2>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. ed., reprint). Psychology Press.
- Cormier, S. M., & Hagman, J. D. (Eds.). (2014). *Transfer of learning: Contemporary research and applications*. Academic Press.
- Davies, M. (2013). Critical thinking and the disciplines reconsidered. *Higher Education Research & Development*, 32(4), 529–544. <https://doi.org/10.1080/07294360.2012.697878>
- Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated reasoning and performance on the Wason selection task. *Personality and Social Psychology Bulletin*, 28(10), 1379–1387. <https://doi.org/10.1177/014616702236869>
- Delaney, P. F., Verhoeven, P. P. J. L., & Spiegel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 53, pp. 63–147). Academic Press. [https://doi.org/10.1016/S0079-7421\(10\)53003-2](https://doi.org/10.1016/S0079-7421(10)53003-2)
- Doll, R. (1982). Clinical trials: Retrospect and prospect. *Statistics in Medicine*, 1(4), 337–344. <https://doi.org/10.1002/sim.4780010411>
- 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(1), 4–58. <https://doi.org/10.1177/1529100612453266>
- Evans, J. S. B. (1977). Toward a statistical theory of reasoning. *Quarterly Journal of Experimental Psychology*, 29(4), 621–635. <https://doi.org/10.1080/14640747708400637>
- Evans, J. S. B. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, 128(6), 978–996. <https://doi.org/10.1037/0033-2909.128.6.978>
- Evans, J. S. B. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10), 454–459. <https://doi.org/10.1016/j.tics.2003.08.012>
- Evans, J. S. B., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11(3), 295–306. <https://doi.org/10.3758/bf03196976>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/brm.41.4.1149>
- Fiorella, L., & Mayer, R. E. (2015). *Learning as a generative activity*. Cambridge University Press. <https://doi.org/10.1017/cbo9781107707085>
- Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, 28(4), 717–741. <https://doi.org/10.1007/s10648-015-9348-9>

- Fitts, D. A. (2010). Improved stopping rules for the design of efficient small-sample experiments in biomedical and biobehavioral research. *Behavior Research Methods*, 42(1), 3–22. <https://doi.org/10.3758/brm.42.1.3>
- Frick, R. W. (1998). A better stopping rule for conventional statistical tests. *Behavior Research Methods, Instruments, & Computers*, 30(4), 690–697. <https://doi.org/10.3758/bf03209488>
- Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15(1), 1–38. [https://doi.org/10.1016/0010-0285\(83\)90002-6](https://doi.org/10.1016/0010-0285(83)90002-6)
- Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(4), 801–812. <https://doi.org/10.1037/a0023219>
- Halpern, D. F. (2014). *Critical thinking across the curriculum: A brief edition of thought & knowledge*. Routledge. <https://doi.org/10.4324/9781315805719>
- Halpern, D. F., & Butler, H. A. (2019). Teaching critical thinking as if our future depends on it, because it does. In J. Dunlosky & K. Rawson (Eds.), *The Cambridge Handbook of Cognition and Education* (pp. 51–66). Cambridge University Press. <https://doi.org/10.1017/9781108235631.004>
- Haskell, R. E. (2001). *Transfer of learning: Cognition, instruction, and reasoning*. Academic Press. <https://doi.org/10.1016/b978-012330595-4/50003-2>
- Heijltjes, A., Van Gog, T., Leppink, J., & Paas, F. (2014). Improving critical thinking: Effects of dispositions and instructions on economics students' reasoning skills. *Learning and Instruction*, 29, 31–42. <https://doi.org/10.1016/j.learninstruc.2013.07.003>
- Heijltjes, A., Van Gog, T., Leppink, J., & Paas, F. (2015). Unraveling the effects of critical thinking instructions, practice, and self-explanation on students' reasoning performance. *Instructional Science*, 43(4), 487–506. <https://doi.org/10.1007/s11251-015-9347-8>
- Heijltjes, A., Van Gog, T., & Paas, F. (2014). Improving students' critical thinking: Empirical support for explicit instructions combined with practice. *Applied Cognitive Psychology*, 28(4), 518–530. <https://doi.org/10.1002/acp.3025>
- Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1989). *Induction: Processes of inference, learning, and discovery*. MIT press.
- Janssen, E. M., Mainhard, T., Buisman, R. S. M., Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Van Peppen, L. M., & Van Gog, T. (2019). Training Higher Education Teachers' Critical Thinking and Attitudes towards Teaching It. *Contemporary Educational Psychology*, 58, 310–322. <https://doi.org/10.1016/j.cedpsych.2019.03.007>
- Janssen, E. M., Meulendijks, W., Mainhard, T., Verkoeijen, P. P., Heijltjes, A. E., Van Peppen, L. M., & Van Gog, T. (2019). Identifying characteristics associated with higher education teachers' Cognitive Reflection Test performance and their attitudes towards teaching critical thinking. *Teaching and Teacher Education*, 84, 139–149. <https://doi.org/10.1016/j.tate.2019.05.008>
- Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review*, 23(1), 1–19. <https://doi.org/10.1007/s10648-010-9150-7>
- Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T., & Camerer, J. F. (2009). The wick in the candle of learning: epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science*, 20(8), 963–973. <http://doi.org/10.1111/j.1467-9280.2009.02402.x>
- Kliegl, O., Bjork, R. A., & Bäuml, K. H. T. (2019). Feedback at test can reverse the retrieval-effort effect. *Frontiers in Psychology*, 10, 1863. <https://doi.org/10.3389/fpsyg.2019.01863>
- Koehler, D. J., Brenner, L., & Griffin, D. (2002). The calibration of expert judgment: Heuristics and biases beyond the laboratory. In T. Gilovich, D. W. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 686–715). Cambridge University Press.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4), 349–370. <https://doi.org/10.1037/0096-3445.126.4.349>
- Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General*, 131(2), 147–162. <https://doi.org/10.1037/0096-3445.131.2.147>
- Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: a distribution-based bifurcation model. *Journal of Memory and Language*, 65(2), 85–97. <https://doi.org/10.1016/j.jml.2011.04.002>
- Kuhn, D. (2005). *Education for thinking*. Harvard University Press.
- Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition*, 17(1), 11–17. <http://doi.org/10.3758/bf03199552>
- Mayer, R. E. (2002). Rote versus meaningful learning. *Theory into Practice*, 41(4), 226–232. https://doi.org/10.1207/s15430421tip4104_4

- 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(3), 360–372. <https://doi.org/10.1002/acp.2914>
- McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory & Cognition*, 1(1), 18–26. <https://doi.org/10.1016/j.jarmac.2011.10.001>
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30–46. <https://doi.org/10.1037/1082-989X.1.1.30>
- Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval practice in classroom settings: a review of applied research. *Frontiers in Education*, 4, 5. <https://doi.org/10.3389/feduc.2019.00005>
- Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science*, 5(4), 207–213. <https://doi.org/10.1111/j.1467-9280.1994.tb00502.x>
- Newstead, S. E., Pollard, P., Evans, J. St. B. T., & Allen, J. L. (1992). The source of belief bias effects in syllogistic reasoning. *Cognition*, 45(3), 257–284. [https://doi.org/10.1016/0010-0277\(92\)90019-e](https://doi.org/10.1016/0010-0277(92)90019-e)
- Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, 84(4), 429–434. <https://doi.org/10.1037/0022-0663.84.4.429>
- Pan, S. C., Gopal, A., & Rickard, T. C. (2016). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology*, 108(4), 563–575. <https://doi.org/10.1037/edu0000074>
- Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: meta-analytic review and synthesis. *Psychological Bulletin*, 144(7), 710–756. <https://doi.org/10.1037/bul0000151>
- Pastötter, B., & Bäuml, K.-H. T. (2016). Reversing the testing effect by feedback: behavioral and electrophysiological evidence. *Cognitive, Affective, & Behavioral Neuroscience*, 16(3), 473–488. <https://doi.org/10.3758/s13415-016-0407-6>
- Perkins, D. N., & Salomon, G. (1992). Transfer of learning. In T. Husen & T. N. Postelwhite (Eds.), *The international encyclopedia of educational* (2nd ed., Vol. 11, pp. 6452–6457). Pergamon Press.
- Pocock, S. J. (1992). When to stop a clinical trial. *British Medical Journal*, 305, 235–240. <https://doi.org/10.1136/bmj.305.6847.235>
- Potts, R., & Shanks, D. R. (2019). The benefit of generating errors during learning: what is the locus of the effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(6), 1023–1041. <https://doi.org/10.1037/xlm0000637>
- Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, 60(4), 437–447. <https://doi.org/10.1016/j.jml.2009.01.004>
- Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, 140(3), 283–302. <https://doi.org/10.1037/a0023956>
- Rawson, K. A., Dunlosky, J., & Sciarrelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review*, 25(4), 523–548. <https://doi.org/10.1007/s10648-013-9240-4>
- Renkl, A. (2014). Toward an instructionally oriented theory of example - based learning. *Cognitive Science*, 38(1), 1–37. <https://doi.org/10.1111/cogs.12086>
- Rickard, T. C., & Pan, S. C. (2017). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review*, 25(3), 847–869. <https://doi.org/10.3758/s13423-017-1298-4>
- Ritchhart, R., & Perkins, D. N. (2005). Learning to think: The challenges of teaching thinking. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 775–802). Cambridge University Press.
- Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1), 20–27. <https://doi.org/10.1016/j.tics.2010.09.003>
- Roediger, H. L., III, & Karpicke, J. D. (2006a). Test enhanced learning: Taking memory tests improves long term retention. *Psychological Science*, 17(3), 249–255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- Roediger, H. L., III, & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1(3), 181–210. <https://doi.org/10.1111/j.1745-6916.2006.00012.x>
- Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(1), 233–239. <https://doi.org/10.1037/a0017678>
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, 140(6), 1432–1463. <https://doi.org/10.1037/a0037559>
- Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanism of a neglected phenomenon. *Educational Psychologist*, 24(2), 113–142. https://doi.org/10.1207/s15326985ep2402_1
- Stanovich, K. E. (2011). *Rationality and the reflective mind*. Oxford University Press.
- Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The Rationality Quotient: Toward a Test of Rational Thinking*. The MIT Press. <https://doi.org/10.7551/mitpress/9780262034845.001.0001>

- Sternberg, Robert J. (2001). Why schools should teach for wisdom: The balance theory of wisdom in educational settings. *Educational Psychologist*, 36(4), 227–245. https://doi.org/10.1207/s15326985ep3604_2
- Storm, B. C., Friedman, M. C., Murayama, K., & Bjork, R. A. (2014). On the transfer of prior tests or study events to subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(1), 115–124. <https://doi.org/10.1037/a0034252>
- Thiede, K. W., Anderson, M. C. M., & Theriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, 95(1), 66–73. <https://doi.org/10.1037/0022-0663.95.1.66>
- Tiruneh, D. T., Verburgh, A., & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. *Higher Education Studies*, 4(1), 1–17. <https://doi.org/10.5539/hes.v4n1p1>
- Tiruneh, D. T., Weldelessie, A. G., Kassa, A., Tefera, Z., De Cock, M., & Elen, J. (2016). Systematic design of a learning environment for domain-specific and domain-general critical thinking skills. *Educational Technology Research and Development*, 64(3), 481–505. <https://doi.org/10.1007/s11423-015-9417-2>
- Trauzettel-Klosinski, S., & Dietz, K. (2012). Standardized assessment of reading performance: the new International Reading Speed Texts IReST. *Investigative Ophthalmology & Visual Science*, 53(9), 5452–5461. <https://doi.org/10.1167/iovs.11-8284>
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- Van Brussel, S., Timmermans, M., Verhoeijen, P., & Paas, F. (2020). 'Consider the Opposite'—Effects of Elaborative Feedback and Correct Answer Feedback on Reducing Confirmation Bias—a Pre-registered Study. *Contemporary Educational Psychology*, 101844. <https://doi.org/10.1016/j.cedpsych.2020.101844>
- Van Gelder, T. (2005). Teaching critical thinking: Some lessons from cognitive science. *College Teaching*, 53(1), 41–48. <https://doi.org/10.3200/ctch.53.1.41-48>
- Van Gog, T., & Paas, F. (2008). Instructional efficiency: revisiting the original construct in educational research. *Educational Psychologist*, 43(1), 16–26. <https://doi.org/10.1080/00461520701756248>
- Van Gog, T., Rummel, N., & Renkl, A. (2019). Learning how to solve problems by studying examples. In J. Dunlosky & K. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 183–208). Cambridge University Press.
- 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(2), 247–264. <https://doi.org/10.1007/s10648-015-9310-x>
- Van Peppen, L. M., Verhoeijen, P. P. J. L., Heijltjes, A. E. G., Janssen, E. M., Koopmans, D., & Van Gog, T. (2018). Effects of self-explaining on learning and transfer of critical thinking skills. *Frontiers in Education*, 3, 100. <https://doi.org/10.3389/feduc.2018.00100>
- Van Peppen, L. M., Verhoeijen, P. P. J. L., Heijltjes, A. E. G., Janssen, E. M., & van Gog, T. (2021). Enhancing students' critical thinking skills: Is comparing correct and erroneous examples beneficial? *Instructional Science*, 1–31. <https://doi.org/10.1007/s11251-021-09559-0>
- Van Peppen, L. M., Verhoeijen, P. P. J. L., Kolenbrander, S. V., Heijltjes, A. E. G., Janssen, E. M., & van Gog, T. (2021). Learning to avoid biased reasoning: Effects of interleaved practice and worked examples. *Journal of Cognitive Psychology*, 33(3), 304–326. <https://doi.org/10.1080/20445911.2021.1890092>
- West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology*, 100(4), 930–941. <https://doi.org/10.1037/a0012842>
- Wittrock, M. C. (2010). Learning as a generative process. *Educational Psychologist*, 45(1), 40–45. <https://doi.org/10.1080/00461520903433554>
- Ximénez, C., & Revuelta, J. (2007). Extending the CLAST sequential rule to one-way ANOVA under group sampling. *Behavior Research Methods, Instruments, & Computers*, 39(1), 86–100. <https://doi.org/10.3758/bf03192847>
- Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25(1), 82–91. <https://doi.org/10.1006/ceps.1999.1016>

SUPPLEMENTARY MATERIALS

Supplemental Material

Download: https://collabra.scholasticahq.com/article/28881-repeated-retrieval-practice-to-foster-students-critical-thinking-skills/attachment/72775.docx?auth_token=U6k9EiuDHJLIbteFEJaS

Peer Review History

Download: https://collabra.scholasticahq.com/article/28881-repeated-retrieval-practice-to-foster-students-critical-thinking-skills/attachment/72776.docx?auth_token=U6k9EiuDHJLIbteFEJaS
