Journal of Cognitive Psychology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/pecp21

Learning to avoid biased reasoning: effects of interleaved practice and worked examples

Lara M. van Peppen, Peter P. J. L. Verkoeijen, Stefan V. Kolenbrander, Anita E. G. Heijltjes, Eva M. Janssen & Tamara van Gog

To cite this article: Lara M. van Peppen, Peter P. J. L. Verkoeijen, Stefan V. Kolenbrander, Anita E. G. Heijltjes, Eva M. Janssen & Tamara van Gog (2021): Learning to avoid biased reasoning: effects of interleaved practice and worked examples, Journal of Cognitive Psychology, DOI: 10.1080/20445911.2021.1890092

To link to this article: https://doi.org/10.1080/20445911.2021.1890092

JOURNAL OF COGNITIVE PSYCHOLOGY
https://doi.org/10.1080/20445911.2021.1890092

Routledge
Taylor & Francis Group

OPEN ACCESS

Learning to avoid biased reasoning: effects of interleaved practice and worked examples

Lara M. van Peppen[a,b], Peter P. J. L. Verkoeijen [a,c], Stefan V. Kolenbrander[c], Anita E. G. Heijltjes[c], Eva M. Janssen[d] and Tamara van Gog[d]

[a]Department of Psychology, Education and Child Studies, Erasmus University Rotterdam Rotterdam, the Netherlands; [b]Institute of Medical Education Research, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands; [c]Learning and Innovation Center, Avans University of Applied Sciences Breda, the Netherlands; [d]Department of Education, Utrecht University Utrecht, the Netherlands

ABSTRACT
It is yet unclear which teaching methods are most effective for improving critical thinking (CT) skills and especially for the ability to avoid biased reasoning. Two experiments (laboratory: $N = 85$; classroom: $N = 117$), investigated the effect of practice schedule (interleaved/blocked) on students' learning and transfer of unbiased reasoning, and whether it interacts with practice-task format (worked-examples/problems). After receiving CT-instructions, participants practiced in: (1) a blocked schedule with worked examples, (2) an interleaved schedule with worked examples, (3) a blocked schedule with problems, or (4) an interleaved schedule with problems. In both experiments, learning outcomes improved after instruction/practice. Surprisingly, there were no indications that interleaved practice led to better learning/transfer than blocked practice, irrespective of task format. The practice-task format did matter for novices' learning: worked examples were more effective than low-assistance practice problems, which demonstrates –for the first time – that the *worked-example effect* also applies to novices' learning to avoid biased reasoning.

ARTICLE HISTORY
Received 5 September 2019
Accepted 9 February 2021

KEYWORDS
Critical thinking; heuristics and biases; contextual interference; interleaved practice; worked examples

Every day, we make many decisions that are based on previous experiences and existing knowledge. This happens almost automatically as we rely on a number of heuristics (i.e. mental shortcuts) that ease reasoning processes (Tversky & Kahneman, 1974). Heuristic reasoning is typically useful, especially in routine situations. But it can also produce systematic deviations from rational norms (i.e. biases; Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1974) with far-reaching consequences, particularly in complex professional environments in which the majority of higher education graduates are employed (e.g. medicine: Ajayi & Okudo, 2016; Elia et al., 2016; Mamede et al., 2010; Law: Koehler et al., 2002). Our primary tool for avoiding bias in reasoning and decision-making (hereafter referred to as *unbiased reasoning;* e.g. Flores et al., 2012; West et al., 2008) is critical thinking (CT). CT-skills are key to effective communication, problem solving, and decision-making in both daily life and professional environments (e.g. Billings & Roberts, 2014; Darling-Hammond, 2010; Kuhn, 2005). Consequently, people who have difficulty with CT are more susceptible to making illogical and biased decisions that can have serious consequences. Given the importance of CT for successful functioning in today's society, it is worrying that many students struggle with several aspects of CT. Hence, it is not surprising that helping students to become critically thinking professionals is a major aim of higher education. However, it is not yet clear what teaching methods are most effective, especially to establish *transfer* (e.g. Van Peppen et al., 2018; Heijltjes et al., 2014a; Heijltjes et al., 2014b, 2015), which refers to the ability to apply acquired knowledge

CONTACT Lara M. van Peppen l.vanpeppen@erasmusmc.nl Institute of Medical Education Research, Erasmus MC, University Medical Center Rotterdam, Doctor Molewaterplein 40, Rotterdam 3051 GD, the Netherlands
Supplemental data for this article can be accessed https://doi.org/10.1080/20445911.2021.1890092

and skills in new situations (Halpern, 1998; Perkins & Salomon, 1992).

Contextual interference in instruction

According to the contextual interference effect, greater transfer is established when materials are presented and learned under conditions of high contextual interference (Schneider et al., 2002). High contextual interference can be created by varying practice-tasks from trial to trial (e.g. Battig, 1978). This task variability induces reflection on to-be-used procedures and can help learners to recognise distinctive characteristics of different problem types (i.e. inter-task comparing) and to develop more elaborate cognitive schemata that contribute to selecting and using a learned procedure when solving similar problems (evidencing learning) and new problems (evidencing transfer; Barreiros et al., 2007; Moxley, 1979).

High contextual interference can be achieved by interleaved practice as opposed to blocked practice. Whereas blocked practice involves practicing one task-category at a time before the next (e.g. AAABBBCCC), interleaved practice mixes practice of several categories together (e.g. ABCBACBCA). To illustrate, a blocked schedule of mathematics tasks first offers practice tasks on volumes of cubes and thereafter practice tasks on volumes of cylinders. An interleaved schedule, on the other hand, offers a mix of practice tasks on volumes of cubes and cylinders. It has been suggested that reflection on the to-be-used procedures is what causes the beneficial effect of interleaved practice (e.g. Barreiros et al., 2007; Rau et al., 2010). Therefore, distinctiveness between task categories should be high enough to reflect what strategy is required, but, on the other hand, should not be too high because learners then immediately recognise what procedure to apply. Additionally, the Sequential Attention Theory (Carvalho & Goldstone, 2019) states that an interleaved schedule highlights differences between items, whereas a blocked schedule highlights similarities between items. Thus, interleaved practice is assumed to be beneficial when differences between categories are crucial for acquiring the category structure. Hence, it is important for beneficial effects of interleaved practice to occur that distinctiveness between categories is high, but distinctiveness within task categories is low (Zulkiply & Burt, 2013). Research on interleaved practice has frequently demonstrated positive learning effects (for a recent meta-analysis, see Brunmair & Richter, 2019), for example in laboratory studies with troubleshooting tasks (De Croock et al., 1998; De Croock & van Merriënboer, 2007; Van Merriënboer et al., 1997, 2002); drawing tasks (Albaret & Thon, 1998); foreign language learning (Abel & Roediger, 2017; Carpenter & Mueller, 2013; Schneider et al., 2002); category induction tasks (Kornell & Bjork, 2008; Sana et al., 2018; Wahlheim et al., 2011); and learning of logical rules (Schneider et al., 1995). Furthermore, several classroom experiments found positive effects of interleaved practice in mathematics learning (e.g. Rau et al., 2013; Rohrer et al., 2014, 2015, 2019), and in astronomy learning (Richland et al., 2005).

The effect of interleaved practice on performance on reasoning tasks has received scant attention in the literature. However, it has been demonstrated with complex judgment tasks that interleaved practice enhanced not only learning but also transfer performance (Helsdingen et al., 2011a, 2011b). In these tasks, participants had to identify relevant cues in case descriptions of, for instance, crimes to estimate priorities of urgency for the police. Although this type of task seems is different from tasks typically used to assess unbiased reasoning (i.e. "heuristics-and-biases tasks"; we will elaborate on these tasks in the materials subsection), both rely on evaluation and interpretation of available information for making appropriate judgments. As such, interleaved practice may have similar effects on learning and transfer of unbiased reasoning.

It is important to note, however, that interleaved practice is usually more cognitively demanding than blocked practice, that is, it places a higher demand on limited working memory resources. Given that it also usually results in better (long-term) learning, interleaved practice seems to impose *germane* cognitive load (Sweller et al., 2011), or "desirable difficulties" (Bjork, 1994). Desirable difficulties are techniques that are effortful during learning and may seem to temporarily hold back performance gains, but are beneficial for long-term performance. Nevertheless, there is a risk that learners, and especially novices, will experience excessively high cognitive load when engaging in interleaved practice, which may hinder learning because it results in the learner being unable to process and compare all relevant information across tasks (Paas & Van Merriënboer, 1994). Using a practice-task format that reduces unnecessary cognitive load, like worked examples (i.e. step-by-step demonstrations of the problem solution; Paas et al., 2003; Renkl, 2014; Sweller, 1988; Van Gog et al., 2019; Van Gog & Rummel, 2010) may help novices benefit from high contextual interference. The high level of guidance during learning from worked examples provides learners with the opportunity to devote attention towards processes – stimulated by interleaved practice – that are directly relevant for learning. As such, learners can use the freed up cognitive capacity to reflect on to-be-used procedures and develop cognitive schemata that contribute to selecting and using a learned procedure when solving similar and novel problems (Kalyuga, 2011; Renkl, 2014). Paas and Van Merriënboer (1994) indeed found that high variability during practice produced transfer test performance benefits (geometrical problem solving) when students studied worked examples, but not when they solved practice problems. Moreover, students who studied worked examples perceived that they invested less mental effort in solving the transfer tasks than did the students who had solved practice problems.

The present study

The aim of the present study was to investigate whether there would be an effect of interleaved practice with heuristics-and-biases tasks on experienced cognitive load, learning outcomes, and transfer performance (e.g. Tversky & Kahneman, 1974) and whether this effect would interact with the format of the practice-tasks (i.e. worked examples or practice problems). We simultaneously conducted 2 experiments: Experiment 1 was conducted in a laboratory setting with university students and Experiment 2 served as a conceptual replication conducted in a real classroom setting with students of a university of applied sciences.[1] Participants received instructions on CT and heuristics and biases tasks, followed by practice with these tasks. Figure 1 displays an overview of the study design: performance was measured as performance on practiced tasks (learning) and non-practiced tasks (transfer), and on a pretest, immediate posttest, and delayed posttest (two weeks later).

In line with previous findings (Van Peppen et al., 2018, submitted; Heijltjes et al., 2014a; Heijltjes et al., 2014b, 2015), we hypothesised that students would benefit from the CT-instructions and practice activities, as evidenced by pretest to immediate posttest gains in performance on practiced items (i.e. *learning*; Hypothesis 1). Regarding our main question (see schematic overview in Table 1), we expected a main effect of interleaved practice, indicating that interleaved practice would require more effort during the practice phase (Hypothesis 2), but would also lead to larger performance gains on practiced items (i.e. *learning*; Hypothesis 3a) and higher performance on non-practiced items (i.e. *transfer*; Hypothesis 3b) than blocked practice. We also expected a main effect of practice-task format: conform the worked example effect, we expected that studying worked examples would be less effortful during the practice phase (Hypothesis 4) and would lead to larger performance gains on practiced items (i.e. *learning*; Hypothesis 5a) and higher performance on non-practiced items (i.e. *transfer*; Hypothesis 5b) than solving problems. Finally, we expected an interaction effect, indicating that the beneficial effect of interleaved practice would be larger with worked examples than practice problems, on both practiced (i.e. *learning*; Hypothesis 6a) and non-practiced (i.e. *transfer*; Hypothesis 6b) items. A delayed (two weeks later) posttest was included, on which we expected these effects (Hypotheses 1-6) to persist. As effects of generative processing (relative to non-generative learning strategies; Dunlosky et al., 2013) and of interleaved practice specifically (Rohrer et al., 2015) sometimes increase as time goes by, they may be even greater after a delay.

Despite not having specific expectations, the mental effort during test data can provide additional insights into the effects of interleaved practice and worked examples on learning (Question 7a/8a) and transfer (Question 7b/8b). As people gain expertise, they can often attain an equal/higher level of performance with less/equal effort investment, respectively. As such, an effort investment decrease in instructed and practiced test items would indicate higher cognitive efficiency (Hoffman & Schraw, 2010; Van Gog & Paas, 2008).[2]

[1] The Dutch education system distinguishes between research-oriented higher education (i.e. offered by research universities) and profession-oriented higher education (i.e. offered by universities of applied sciences).

[2] We also exploratively analyzed students' global judgments of learning (JOLs) after practice to gain insight into how informative the different practice types were according to the students themselves; however, these analyses did not have much added value for this paper, and, therefore, are not reported here but provided on our OSF-page.

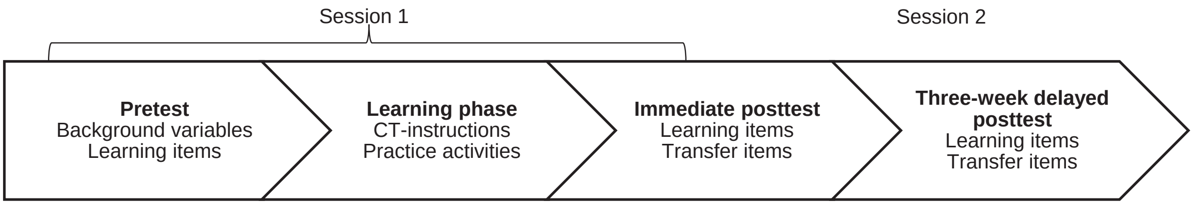

Figure 1. Overview of the study design. The four conditions differed in practice activities during the learning phase.

Table 1. Schematic overview of hypotheses 2–6.

	Mental effort during learning	**Test performance**	
		Learning items	Transfer items
Practice schedule	Interleaved > Blocked (hypothesis 2)	Interleaved > Blocked (hypothesis 3a)	Interleaved > Blocked (hypothesis 3b)
Practice-task format	Examples < Problems (hypothesis 4)	Examples > Problems (hypothesis 5a)	Examples > Problems (hypothesis 5b)
Interaction Practice schedule and Practice-task format		Effect Interleaved over Blocked: Examples > Problems (hypothesis 6a)	Effect Interleaved over Blocked: Examples > Problems (hypothesis 6b)

Note: Additional research questions were formulated regarding the mental effort invested in the test (Question 7 and 8), but these are not provided in this table because we did not have specific expectations.

Experiment 1

Materials and methods

We created an Open Science Framework (OSF) page for this project, where detailed descriptions of the experimental design and procedures are provided and where all data and materials (in Dutch) can be found (osf.io/a9czu).

Participants

Participants were 112 first-year Psychology students of a Dutch university. Of these, 104 students (93%) were present at both experimental sessions (see the procedure subsection for more information), and only their data were analysed. Participants were excluded from the analyses when test or practice sessions were not completed or when instructions were not adhered to, i.e. when more than half of the practice tasks were not read seriously. Based on the fact that fast readers can read no more than 350 words per minute (e.g. Trauzettel-Klosinski & Dietz, 2012) – and the words in these tasks additionally require understanding – we assumed that participants who spent less than 0.17 s per word (i.e. 60 s/350 words) did not read the instructions seriously. This involved more participants from the worked examples conditions than the practice problems conditions and resulted in a final sample of 85 students ($M_{\text{age}} = 19.84$, SD = 2.41; 14 males). Based on this sample size, we have calculated a power function of our analyses using the G*Power software (Faul et al., 2009). The power of Experiment 1 – under a fixed alpha level of 0.05 and with a correlation between measures of 0.3 (e.g. Van Peppen et al., 2018) – is estimated at .24 for detecting a small interaction effect ($\eta_p^2 = .01$), .96 for a medium interaction effect ($\eta_p^2 = .06$), and > .99 for a large interaction effect ($\eta_p^2 = .14$). Thus, the power of our experiment should be sufficient to pick up medium-sized interaction effects, which is in line with the moderate overall positive effect of interleaved practice of previous studies as indicated in a recent meta-analysis ($g = 0.42$; Brunmair & Richter, 2019).[3]

Design

The experiment consisted of four phases (see Figure 1): pretest, learning phase (CT-instructions plus practice), immediate posttest, and delayed posttest. A 3 × 2 × 2 design was used, with Test Moment (pretest, immediate posttest, and delayed posttest) as within-subjects factor and Practice Schedule

[3] In response to a reviewer, we have calculated power functions of our post hoc analyses. The power of the comparison between interleaved practice and blocked practice, under a fixed alpha level of 0.05, is estimated at .15, .62, and .95 for detecting a small ($d = .02$), medium ($d = .05$), and large ($d = .08$) effect, respectively. The power of the comparison between worked examples and practice problems is estimated at .15, .60, and .95 for detecting a small, medium, and large effect, respectively. Thus, the power of our experiment should be sufficient to pick up medium-to-large-sized effects. However, the power to pick up a differential effect of interleaved practice with worked examples compared to practice problems seems relatively low, to wit, .09, .33, and .67 for detection of a small, medium, or large effect, respectively.

(interleaved and blocked) and Practice-task Format (worked examples and practice problems) as between-subjects factors. After completing the pretest on learning items (i.e. instructed and practiced during the learning phase), participants received instructions and were randomly assigned to one of four practice conditions: (1) Blocked Schedule with Worked Examples Condition ($n = 18$); (2) Blocked Schedule with Practice Problems Condition ($n = 28$); (3) Interleaved Schedule with Worked Examples Condition ($n = 17$); and (4) Interleaved Schedule with Practice Problems Condition ($n = 22$). Subsequently, participants completed the immediate posttest and two weeks later the delayed posttest on learning items (i.e. instructed and practiced during the learning phase) and transfer items (i.e. not instructed and practiced during the learning phase).

Materials

All materials were delivered in a computer-based environment (Qualtrics platform) that is created for this study.

CT-skills tests. The CT-skills pretest consisted of nine classic heuristics-and-biases items across three categories (e.g. West et al., 2008) which we refer to as *learning items* as (isomorphs of) these items were instructed and practiced during the learning phase, (example-items in Appendix): (1) Base-rate items which measured the tendency to overweigh individual-case evidence, that is, specific information (e.g. from personal experience, a single case, or prior beliefs) and to undervalue statistical information (Stanovich et al., 2016; Stanovich & West, 2000; Tversky & Kahneman, 1974); (2) Conjunction items that measured to what extent the conjunction rule ($P(A\&B) \leq P(B)$) is neglected – this fundamental rule in probability theory states that the probability of Event A and Event B both occurring must be lower than the probability of Event A or Event B occurring alone (adapted from Tversky & Kahneman, 1983); (3) Syllogistic reasoning items that examined the tendency to be influenced by the believability of a conclusion when evaluating the logical validity of arguments (Evans, 2003). As mentioned previously, it is important for interleaved practice effects to occur that distinctiveness between categories is high enough to reflect what strategy is required but, on the other hand, is not too high because learners then immediately recognise what procedure to apply (see for example, Brunmair & Richter, 2019; Carvalho & Goldstone, 2019). Therefore, we combined lower distinctive task categories (i.e. only requiring knowledge and rules of *statistics*: base-rate vs. conjunction) with higher distinctive task categories (i.e. requiring knowledge and rules of *statistics and logic*: base-rate vs. syllogistic reasoning and conjunction vs. syllogistic reasoning).

The immediate and delayed posttest contained parallel versions of the nine pretest learning items across three categories (base-rate, conjunction, and syllogism) that were designed as structurally equivalent but with different surface features. To illustrate, an immediate posttest item contained the exact same wording as the respective pretest item but, for instance, described a different company. In addition, the immediate and delayed posttests also contained four items of two task-categories that were *transfer items* as these were not instructed and practiced during the learning phase. The transfer items shared similar features with the learning items, namely, requiring knowledge and rules of logic (i.e. syllogisms rules) or requiring knowledge and rules of statistics (i.e. probability and data interpretation), respectively: (1) Wason selection items which measured the tendency to confirm a hypothesis rather than to falsify it (adapted from Evans, 2002; Gigerenzer & Hug, 1992); and (2) Contingency items measured the tendency to judge information given in a contingency table unequally, based on already experienced evidence (Heijltjes et al., 2014a; Stanovich & West, 2000; Wasserman et al., 1990).

In the interleaved schedule, all items were offered in random order and in the blocked schedule the items were randomly offered within the blocks. A multiple-choice (MC) format with different numbers of alternatives per item was used, with only one correct alternative for each task that evidences unbiased reasoning. The incorrect alternatives were intuitive (and incorrect) responses or results of incomplete reasoning processes. The content of the surface features (cover stories) of all test items was adapted to the study domain of the participants. All conditions were pilot-tested on difficulty, duration, and representativeness of content (for the study programme) by some students from a university of applied sciences (not partaking in the main experiments). Moreover, several tasks were taken from previous studies that were conducted in similar contexts (i.e. within an existing CT-course with first-year or second-year students of a university of applied sciences; Heijltjes et al., 2014a; Heijltjes et al., 2014b, 2015) and even within the same study domain (Van Peppen et al., 2018).

CT-instructions. The video-based instruction consisted of a general instruction on CT and explicit instructions on three heuristics-and-biases tasks. In the general instruction, the features of CT and the attitudes and skills that are needed to think critically were described. Thereafter, participants received explicit instructions on how to avoid base-rate fallacies, conjunction fallacies, and biases in syllogistic reasoning. These instructions consisted of a worked example of each category that not only showed the correct line of reasoning but also included possible problem-solving strategies. The worked examples provided solutions to the tasks seen in the pretest, which allowed participants to mentally correct initially erroneous responses.

CT-practice. The CT-practice phase consisted of nine practice tasks across the three task categories – in random order – of the pretest and the explicit instructions: base-rate (Br), conjunction (C), and syllogistic reasoning (S). Depending on the assigned condition, participants had to practice either in an interleaved (e.g. Br–C-S–C-S-Br-S-Br–C) or blocked schedule (e.g. Br-Br-Br–C-C–C-S-S-S), and either with worked examples or practice problems. Participants in the practice problems conditions were instructed to read the tasks thoroughly and to choose the best answer option. They received a prompt after each of the tasks in which they were asked to explain how the answer was obtained. After that, participants received feedback indicating whether the given answer was correct or incorrect (i.e. "your answer to this assignment was correct" or "your answer to this assignment was incorrect"). Participants in the worked examples conditions were first told that they would not have to solve the problems themselves, but that they receive a worked-out solution to each problem. They were instructed to read each worked-out example thoroughly. The worked examples consisted of a problem statement and a solution to this problem (i.e. the strategy information provided during the CT-instructions was repeated in the worked examples). The line of reasoning and underlying principles were explained in steps, sometimes clarified with a visual representation. The explanations given in the worked examples were based on the explanations from the original literature on the tasks (e.g. "to solve this problem you should … ") and have been rewritten to make it look like another student has completed the task (e.g. "to solve this problem, I am … "). Thus, the worked examples consisted of more elaborate information compared to the practice problems.

Mental effort. Invested mental effort was measured with the subjective rating scale developed by Paas (1992). After each practice-task and after each test item, participants reported how much mental effort they invested in completing that task or item, on a 9-point scale ranging from (1) very, very low effort to (9) very, very high effort.

Procedure

The study was run in two sessions that both took place in the computer lab of the university. Participants signed an informed consent form at the start of the experiment. Before participants arrived, A4-papers were distributed among all cubicles (one participant in each cubicle) containing some general rules and a link to the Qualtrics environment of session 1, where all materials were delivered. Participants could work at their own pace and time-on-task was logged during all phases. Furthermore, participants were allowed to use scrap paper during the practice phase and the CT-tests.

In session 1 (ca. 75 min), participants first filled out a demographic questionnaire and then completed the pretest. After each test item, they had to indicate how much mental effort they invested in it. Subsequently, participants entered the learning phase in which they first viewed the video (10 min.), including the general CT-instruction and the explicit instructions. Thereafter, the Qualtrics programme randomly assigned the participants to one of the four practice conditions. Participants rated after each practice task how much mental effort they invested. After the learning phase, participants completed the immediate posttest and again rated their invested mental effort after each test item. The second session took place two weeks later and lasted circa 20 min. Participants again received an A4-paper containing some general rules and a link to the Qualtrics environment of session 2. This time, participants completed the delayed posttest and again reported their mental effort ratings after each test item. One experiment leader (first or third author of this paper) was present during all phases of the experiment.

Data analysis. Of the nine learning items of the CT-skills tests, seven items were MC-only questions (with more than two alternatives) and two items were MC-plus-motivation questions (with two MC alternatives; one conjunction and one base-rate item) to prevent participants from guessing. The transfer items consisted of two MC-only and two MC-plus-motivation questions (two contingency items). Performance on the pretest, immediate posttest, and delayed posttest was scored by assigning 1 point to each correct alternative on the MC-only questions (i.e. referring to unbiased reasoning). For items with only two MC alternatives, the scoring was based on the explanation provided so that no points were assigned for correct guesses. Participants could earn 1 point for the correct explanation, 0.5 point for a partially correct explanation,[4] and 0 points for an incorrect explanation for these MC-plus-motivation questions (score form developed by the first author). As a result, participants could earn a maximum score of 9 on the learning items and a maximum total score of 4 on the transfer items. Two raters independently scored 25% of the explanations on the open questions of the immediate posttest, blind to student identity and condition. The intra-class correlation coefficient was .991 for the learning test items and .986 for the transfer test items. Because of the high inter-rater reliability, the remainder of the tests was scored by one rater (the first author) and this rater's scores were used in the analyses..

For comparability, we computed percentage scores on the learning and transfer items instead of total scores. It is important to realise that, even though we used percentage scores, caution is warranted in interpreting differences between learning and transfer outcomes because the maximum scores differed. The mean score on the posttest learning items was 59.9% (SD = 20.22) and reliability of these items (Cronbach's alpha) was .24 on the pretest, .57 on the immediate posttest, and .51 on the delayed posttest. The low reliability on the pretest might be explained by the fact that a lack of prior knowledge requires guessing of answers. As such, inter-item correlations are low, resulting in a low Cronbach's alpha. Moreover, caution is required in interpreting these reliabilities because sample sizes as in studies like this do not seem to produce sufficiently precise alpha coefficients (e.g. Charter, 2003). The mean score on the posttest transfer items was 36.2% (SD = 22.31). Reliability of these items was low (Cronbach's alpha of .25 on the posttest and .43 on the delayed posttest), which can probably partly be explained by floor effects at both tests for one of our transfer task categories (i.e. Wason selection). Therefore, we decided not to report the test statistics of the analyses on transfer performance. Descriptive statistics can be found in Tables 2 and 3.

Results

In all analyses reported below, a significance level of .05 was used. Partial eta-squared (η_p^2) is reported as a measure of effect size for the ANOVAs for which 0.01 is considered small, 0.06 medium, and 0.14 large (Cohen, 1988). On our OSF-project page we presented the intention-to-treat (i.e. all participants who entered the study) analyses, which did not reveal noteworthy differences with the compliant-only (i.e. all participants who have met the criterion of spending more than 0.17 s per word for at least half of the practice tasks) analyses reported below.

Check on condition equivalence and time-on-task

Following the drop-out of some participants, we checked our conditions on equivalence. Preliminary analyses confirmed that the conditions did not differ in educational background, $\chi^2(15) = 15.68$, $p = .403$; performance on the pretest, $F(3, 81) = 1.68$, $p = .178$; time spent on the pretest, $F(3, 81) = 1.75$, $p = .164$; and average mental effort invested on the pretest items, $F(3, 81) = 0.78$, $p = .510$. We found a gender difference between the conditions, $\chi^2(3) = 11.03$, $p = .012$. However, gender did not correlate significantly with learning performance (minimum $p = .108$) and was therefore not a confounding variable.

A 2 (Practice Schedule: interleaved vs. blocked) × 2 (Practice-task Format: worked examples vs. practice problems) factorial ANOVA showed no significant differences on time-on-task during practice between the interleaved and blocked conditions, $F(3, 81) = 3.05$, $p = .085$, $\eta_p^2 = .04$, but there was a significant difference between worked examples conditions ($M = 577.48$, $SE = 37.93$) compared to the practice problems conditions ($M = 737.61$, $SE = 31.96$), $F(3, 81) =$

[4] That is, when half of the necessary information was given. To illustrate, a correct explanation on a contingency table involves correct consideration of the information presented in the rows and columns, while a partially correct explanation only involves consideration of either the information in the rows or the information in the columns.

Table 2. Means (SD) of Test performance (multiple-choice % score) and Invested Mental Effort (1-9) per Condition of Experiment 1.

		Instructional conditions			
		Blocked Schedule Worked Examples	Blocked Schedule Practice Problems	Interleaved Schedule Worked Examples	Interleaved Schedule Practice Problems
Test performance					
Learning items	Pretest	23.46 (13.14)	29.37 (13.60)	24.18 (11.94)	20.20 (13.56)
	Immediate posttest	65.43 (23.15)	55.95 (18.27)	71.90 (18.89)	51.01 (15.96)
	Delayed posttest	68.86 (19.53)	59.13 (17.12)	73.86 (17.98)	53.54 (15.58)
Transfer items	Immediate posttest	43.06 (22.37)	40.63 (22.21)	36.03 (19.71)	26.70 (22.26)
	Delayed posttest	47.22 (24.08)	45.54 (18.07)	39.71 (28.03)	50.00 (18.90)
Mental effort during test					
Learning items	Pretest	3.47 (0.99)	3.73 (0.66)	3.84 (0.63)	3.76 (0.89)
	Immediate posttest	3.28 (1.23)	3.97 (0.99)	3.80 (0.58)	3.80 (0.90)
	Delayed posttest	3.25 (1.01)	4.09 (0.97)	3.80 (0.88)	4.20 (0.88)
Transfer items	Immediate posttest	4.14 (1.38)	4.81 (1.10)	4.85 (0.72)	4.81 (0.97)
	Delayed posttest	3.81 (1.45)	4.57 (0.80)	4.46 (0.98)	5.01 (0.94)
Mental effort during learning		3.51 (0.26)	4.05 (0.21)	4.20 (0.26)	4.11 (0.23)

Table 3. Means (SD) of Test performance per task (max. score 1) per Condition of Experiment 1.

		Instructional conditions			
		Blocked Examples	Blocked Problems	Interleaved Examples	Interleaved Problems
Syllogism 1	Pretest	0.67 (0.49)	0.75 (0.44)	0.53 (0.51)	0.55 (0.51)
	Immediate posttest	0.50 (0.51)	0.43 (0.50)	0.47 (0.51)	0.55 (0.51)
	Delayed posttest	0.78 (0.43)	0.54 (0.51)	0.65 (0.49)	0.64 (0.49)
Syllogism 2	Pretest	0.06 (0.24)	0.14 (0.36)	0.00 (0.00)	0.09 (0.29)
	Immediate posttest	0.61 (0.50)	0.64 (0.49)	0.71 (0.47)	0.55 (0.51)
	Delayed posttest	0.39 (0.50)	0.39 (0.50)	0.47 (0.51)	0.27 (0.46)
Syllogism 3	Pretest	0.17 (0.38)	0.18 (0.39)	0.00 (0.00)	0.14 (0.35)
	Immediate posttest	0.33 (0.49)	0.18 (0.39)	0.71 (0.47)	0.09 (0.29)
	Delayed posttest	0.56 (0.51)	0.64 (0.49)	0.71 (0.47)	0.55 (0.51)
Base-rate 1	Pretest	0.00 (0.00)	0.04 (0.19)	0.00 (0.00)	0.00 (0.00)
	Immediate posttest	0.56 (0.51)	0.46 (0.51)	0.65 (0.49)	0.36 (0.49)
	Delayed posttest	0.44 (0.51)	0.50 (0.51)	0.71 (0.47)	0.27 (0.46)
Base-rate 2	Pretest	0.06 (0.27)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Immediate posttest	0.44 (0.51)	0.04 (0.19)	0.24 (0.44)	0.00 (0.00)
	Delayed posttest	0.28 (0.46)	0.00 (0.00)	0.24 (0.44)	0.00 (0.00)
Base-rate 3	Pretest	0.67 (0.49)	0.79 (0.42)	0.82 (0.39)	0.59 (0.50)
	Immediate posttest	0.89 (0.32)	0.79 (0.42)	1.00 (0.00)	0.68 (0.48)
	Delayed posttest	1.00 (0.00)	0.75 (0.44)	1.00 (0.00)	0.68 (0.48)
Conjunction 1	Pretest	0.11 (0.32)	0.14 (0.36)	0.24 (0.44)	0.18 (0.39)
	Immediate posttest	0.78 (0.43)	0.86 (0.36)	0.88 (0.33)	0.73 (0.46)
	Delayed posttest	0.89 (0.32)	0.89 (0.32)	0.88 (0.33)	0.77 (0.43)
Conjunction 2	Pretest	0.22 (0.43)	0.36 (0.49)	0.29 (0.47)	0.18 (0.40)
	Immediate posttest	0.83 (0.38)	0.79 (0.42)	0.94 (0.24)	0.77 (0.43)
	Delayed	0.94 (0.24)	0.75 (0.44)	1.00 (0.00)	0.82 (0.40)
Conjunction 3	Pretest	0.17 (0.38)	0.25 (0.44)	0.29 (0.47)	0.09 (0.29)
	Immediate posttest	0.94 (0.24)	0.86 (.36)	0.88 (0.33)	0.86 (0.35)
	Delayed posttest	0.89 (0.32)	0.86 (.36)	1.00 (0.00)	0.82 (0.39)
Wason selection 1	Immediate posttest	0.11 (0.32)	0.11 (.32)	0.00 (0.00)	0.14 (0.35)
	Delayed posttest	0.06 (0.24)	0.00 (.00)	0.00 (0.00)	0.09 (0.29)
Wason selection 2	Immediate posttest	0.17 (0.38)	0.29 (.46)	0.12 (0.33)	0.14 (0.35)
	Delayed posttest	.28 (0.46)	0.18 (.39)	0.29 (0.47)	0.14 (0.35)
Contingency 1	Immediate posttest	0.69 (0.42)	0.61 (.48)	0.65 (0.42)	0.34 (0.42)
	Delayed posttest	0.72 (0.46)	0.75 (.42)	0.56 (0.50)	0.68 (0.42)
Contingency 2	Immediate posttest	0.72 (0.43)	0.54 (0.47)	0.68 (0.47)	0.32 (0.45)
	Delayed posttest	0.69 (0.42)	0.75 (0.40)	0.59 (0.48)	0.77 (0.34)

10.42, $p = .002$, $\eta_p^2 = .11$. If it turns out that the practice problems conditions outperformed the worked examples conditions, this finding should be taken into account. No significant interaction between Practice Schedule and Practice-task Format was found, $F(3, 81) = 1.00$, $p = .320$, $\eta_p^2 = .01$.[5]

[5] The relatively low reliabilities of the learning items should be taken into account.

Performance on learning items

Performance data are presented in Tables 2 and 3 and all omnibus test statistics can be found in Table 4 (statistics of follow-up analyses are presented in text). A 3 × 2 × 2 mixed ANOVA on the items that assessed learning, with Test Moment (pretest, immediate posttest, and delayed posttest) as within-subjects factor and Practice Schedule (interleaved and blocked) and Practice-task Format (worked examples and practice problems) as between-subjects factors, showed a main effect of Test Moment. In line with Hypothesis 1, repeated contrasts revealed that participants performed better on the immediate posttest ($M = 61.07$, $SE = 2.10$) than on the pretest ($M = 24.30$, $SE = 1.46$), $F(1, 81) = 267.66$, $p < .001$, $\eta_p^2 = .77$. There was no significant difference between performance on the immediate and delayed posttest ($M = 63.76$, $SE = 1.93$), $F(1, 81) = 2.90$, $p = .092$, $\eta_p^2 = .04$.

In contrast to Hypothesis 3a (see Table 1 for a schematic overview of the hypotheses), we did not find a significant main effect of Practice Schedule or an interaction between Practice Schedule and Test Moment on performance on learning items. However, the analysis did reveal a main effect of Practice-task Format, with worked examples resulting in better performance ($M = 54.56$, $SE = 2.21$) than practice problems ($M = 44.87$, $SE = 1.86$). This was qualified by an interaction effect between Practice-task Format and Test Moment: in line with Hypothesis 5a, repeated contrasts revealed that there was a higher pretest to immediate posttest performance gain for worked examples ($M_{\text{pre}} = 23.82$, $SE = 2.23$; $M_{\text{immediate}} = 68.66$, $SE = 3.21$) than for practice problems ($M_{\text{pre}} = 24.78$, $SE = 1.88$; $M_{\text{immediate}} = 53.48$, $SE = 2.70$), $F(1, 81) = 12.90$, $p = .001$, $\eta_p^2 = .14$. Contrary to Hypothesis 6a, there was no interaction between Practice Schedule and Practice-task Format, nor an interaction between Practice Schedule, Practice-task Format, and Test Moment.

Mental effort during learning

Mental effort data are presented in Table 2 and all omnibus test statistics can be found in Table 4. Contrary to hypotheses 2 and 4 respectively, a 2 (Practice Schedule: interleaved and blocked) × 2 (Practice-task Format: worked examples and practice problems) factorial ANOVA on the mental effort during practice data revealed no main effects of Practice Schedule and Practice-task

Table 4. Results Mixed ANOVAs Experiment 1.

		Test performance			Mental effort		
	ANOVA	F-test (df)	p*	η_p^2	F-test (df)	p*	η_p^2
Learning items	Test Moment	242.29 (2,162)	< .001*	.75	1.15 (1.837,148.825)	.315	.01
	Test Moment × Practice Schedule	0.88 (2,162)	.417	.01	0.35 (1.837,148.825)	.689	.00
	Test Moment × Practice-task Format	10.62 (2,162)	< .001*	.12	3.55 (1.837,148.825)	.035*	.04
	Test Moment × Practice Schedule × Practice-task Format	0.01 (2,162)	.981	.00	0.40 (1.837,148.825)	.654	.01
	Practice Schedule	0.17 (1,81)	.680	.00	2.11 (1,81)	.150	.03
	Practice-task Format	11.30 (1,81)	.001*	.12	4.74 (1,81)	.032*	.06
	Practice Schedule × Practice-task Format	3.47 (1,81)	.066	.04	2.28 (1,81)	.135	.03
Practice tasks	Practice Schedule	–	–	–	2.41 (1,81)	.125	.03
	Practice-task Format	–	–	–	0.88 (1,81)	.352	.01
	Practice Schedule × Practice-task Format	–	–	–	1.72 (1,81)	.194	.02

*$p < .05$.

Format. Moreover, no interaction between Practice Schedule and Practice-task Format was found.

Mental effort during test

We exploratory analysed the mental effort during test data with a $3 \times 2 \times 2$ mixed ANOVA on mental effort invested on learning items and a $2 \times 2 \times 2$ mixed ANOVA on mental effort invested on transfer items (i.e. transfer items were not included in the pretest). Mental effort data during test is presented in Table 2 and all test statistics can be found in Table 4.

Regarding effort invested in the *learning items*, there was no main effect of Practice Schedule (Question 7a). However, there was a main effect of Practice-task Format (Question 8a); less invested effort on learning items was reported in the worked examples conditions ($M = 3.57$, $SE = .13$) compared to practice problems conditions ($M = 3.92$, $SE = .11$), and an interaction effect between Test Moment and Practice-task Format. Repeated contrasts revealed an effort investment increase over time with a significant difference between immediate and delayed posttest for the practice problems conditions ($M_{\text{pretest}} = 3.74$, $SE = .11$; $M_{\text{immediate}} = 3.89$, $SE = .14$; $M_{\text{delayed}} = 4.14$, $SE = .13$), $F(1,48) = 6.08$, $p = .017$, $\eta_p^2 = .11$, and no significant differences for the worked examples conditions, $F(2,66) = .38$, $p = .683$, $\eta_p^2 = .01$. The results did not reveal a main effect of Test Moment and interaction effects.

Regarding invested mental effort in the *transfer items,* the results revealed a main effect of Practice Schedule (Question 7b), with higher effort investment when practiced in an interleaved schedule ($M = 4.78$, $SD = .15$) compared to a blocked schedule ($M = 4.33$, $SD = .14$). Furthermore, there was an effect of Practice-task Format (Question 7b): higher effort investment was reported by the practice problems conditions ($M = 4.80$, $SD = .13$) compared to worked examples conditions ($M = 4.31$, $SD = .16$). No main effect of Test Moment and interaction effects were found.

Interim summary

Taken together, there were no indications that interleaved practice – either in itself or as a function of task-format – contributed to better learning. However, interleaved practice resulted in higher effort investment on transfer items than blocked practice, which may indicate that interleaved practice stimulated analytical and effortful reasoning (i.e. Type 2 processing, e.g. Stanovich, 2011) more than blocked practice yet without resulting in replacement of the incorrect intuitive response (i.e. Type 1 processing) with the more analytical correct response. Alternatively, this finding may indicate a lower cognitive efficiency (Hoffman & Schraw, 2010; Van Gog & Paas, 2008) of interleaved practice as opposed to blocked practice. Furthermore, in line with the worked example effect (e.g. Sweller et al., 2011), studying worked examples was more effective for learning than solving problems, as well as more efficient (i.e. higher test performance reached in less practice time and less mental effort investment during the test phase; Van Gog & Paas, 2008). We will further elaborate on and discuss the findings of Experiment 1 in the General Discussion.

Experiment 2

We simultaneously conducted a replication experiment in a classroom setting to assess the robustness of our findings and to increase ecological validity. All test and practice items were the same but, if necessary, adapted to the domain of the participants to meet the requirements of the study programme (see for example the conjunction item in the appendix).

Materials and methods

Participants and design

The design of Experiment 2 was the same as that of Experiment 1. Participants were 157 second-year "Safety and Security Management" students of two locations of a Dutch university of applied sciences. Students from the first location had some prior knowledge as they had participated in a study that included similar heuristics-and-biases tasks in the first year of their curriculum that was followed by some lessons on this topic ($n = 83$), while students of the second location ($n = 74$) had not. Since the level of prior knowledge may be relevant (Likourezos et al., 2019), the factor Site will be included in the main analyses. Of the 157, 117 students (75%) were present at both sessions. As a large number of students missed the second session, we decided to conduct two separate analyses on performance and mental effort on learning items (transfer items were only included in the immediate and delayed posttest): pretest to immediate posttest analyses for all students present during session 1 and immediate posttest to delayed posttest analyses for all students present at both sessions. As in Experiment 1, participants who did not read the instructions seriously were excluded of the analyses. This resulted in a final subsample of 117 students ($M_{\text{age}} = 20.05$, SD = 1.76; 70 males; 60 higher knowledge) for the pretest-immediate posttest analyses and a final subsample of 89 students ($M_{\text{age}} = 19.92$, SD = 1.78; 46 males; 51 higher-knowledge) for the immediate posttest-delayed posttest analyses. Participants were randomly assigned to the Blocked Schedule with Worked Examples ($n = 20$; $n = 15$); Blocked Schedule with Practice Problems ($n = 43$; $n = 33$); Interleaved Schedule with Worked Examples ($n = 15$; $n = 8$); and Interleaved Schedule with Practice Problems ($n = 39$; $n = 32$) conditions. Based on these two sample sizes, we have calculated power functions of Experiment 2 using the G*Power software (Faul et al., 2009), including the factor Site. The power of analysis 1 ($n = 117$) – under a fixed alpha level of 0.05 and with a correlation between measures of 0.3 (e.g. Van Peppen et al., 2018) – is estimated at .20 for detecting a small interaction effect ($\eta_p^2 = .01$), .93 for a medium interaction effect ($\eta_p^2 = .06$), and > .99 for a large interaction effect ($\eta_p^2 = .14$). Under the same assumptions, the power of analysis 2 ($n = 89$) is estimated at .17 for detecting a small interaction effect ($\eta_p^2 = .01$), .82 for a medium interaction effect ($\eta_p^2 = .06$), and > .99 for a large interaction effect ($\eta_p^2 = .14$). Thus, our experiment should be sufficient to pick up medium-sized interaction effects, which could be expected given the moderate overall positive effect of interleaved practice found in previous studies (Brunmair & Richter, 2019).[6]

Materials, procedure, and scoring

All data, materials, and detailed descriptions of the procedures and scoring are provided at the OSF-page of this project. The same materials were used as in Experiment 1 but the content of the surface features (cover stories) was adapted to the domain of the participants when the original features did not reflect realistic situations for these participants to keep the level of difficulty approximately equal to Experiment 1 and to meet the requirements of the study programme (i.e. the final exam was based on these materials). The content of all materials was evaluated, including equivalence of information, and approved by a teacher working in the domain.

The main difference with Experiment 1 was that Experiment 2 was run in a real education setting, namely during the lessons of a CT-course. Experiment 2 was conducted in a computer classroom at the participants' school with an entire class of students present. Participants came from eight different classes (of 25–31 participants) and were randomly distributed among the four conditions within each class. The two sessions of Experiment 2 took place during the first two lessons and between these lessons no CT- instruction was given. In advance of the first session, students were informed about the experiment by their teacher. When entering the classroom, 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 of session 1 where they first signed an informed consent form. Again, participants could work at their own pace and could use scrap paper and time-on-task was logged during all phases. Participants had to wait (in silence) until the last participant had finished the posttest before they were allowed to leave the classroom. The experiment leader and the teacher of the CT-course (first and third author of this paper) were both present during all phases of the experiment and one of them explained the nature of the experiment afterwards.

The same test-items and score form for the open questions were used as in Experiment 1. Again, participants could attain a maximum score of 9 on the learning items and a maximum total score of 4 on the transfer items and we computed percentage scores on the learning and transfer items instead of total scores. It is important to realise that, even though we used percentage scores, caution is warranted in interpreting differences between learning and transfer outcomes because the maximum scores differed. Two raters independently scored

[6] In response to a reviewer, we calculated power functions of our post hoc analyses. The power of the comparison between interleaved practice and blocked practice, under a fixed alpha level of 0.05, is estimated at .19, .76, and >.99 (analysis 1) and .15, .64, and .96 (analysis 2) for detecting a small (d = .02), medium (d = .05), and large (d = .08) effect, respectively. The power of the comparison between worked examples and practice problems is estimated at .17, .69, and .98 (analysis 1) and .13, .53, and .90 (analysis 2), for detecting a small, medium, and large effect, respectively. Thus, the power of our experiment should be sufficient to pick up medium-to-large-sized effects of interleaved practice vs. blocked practice and large-sized effects of worked examples vs. practice problems. However, the power to pick up a differential effect of interleaved practice with worked examples compared to practice problems seems relatively low, to wit, .10, .37, and .73 (analysis 1) and .08, .23, and .50 (analysis 2) for detection of a small, medium, or large effect, respectively.

Table 5. Means (SD) of Test performance (multiple-choice % score) and Invested Mental Effort (1-9) per condition and analysis of Experiment 2.

		Instructional conditions			
		Blocked Schedule Worked Examples	Blocked Schedule Practice Problems	Interleaved Schedule Worked Examples	Interleaved Schedule Practice Problems
Analysis 1					
Test performance					
Learning items	Pretest	35.56 (20.58)	41.09 (20.65)	40.00 (20.91)	43.59 (27.55)
	Immediate posttest	68.33 (15.83)	56.85 (21.17)	75.56 (15.83)	60.68 (16.49)
Mental effort during test					
Learning items	Pretest	3.81 (0.99)	4.01 (.87)	3.97 (1.09)	4.23 (1.08)
	Immediate posttest	3.78 (1.10)	3.86 (1.09)	3.78 (1.10)	4.36 (0.95)
Analysis 2					
Test Performance					
Learning items	Immediate posttest	68.15 (16.19)	58.25 (21.70)	72.22 (18.78)	62.50 (14.87)
	Delayed posttest	71.85 (16.19)	63.64 (22.95)	70.83 (19.64)	70.14 (13.37)
Transfer items	Immediate posttest	30.83 (22.04)	27.65 (22.04)	26.39 (19.21)	30.86 (26.56)
	Delayed posttest	35.83 (19.97)	32.20 (21.43)	33.33 (20.84)	28.13 (22.67)
Mental effort during test					
Learning items	Immediate posttest	3.80 (1.11)	3.83 (0.99)	3.65 (1.65)	4.42 (0.97)
	Delayed posttest	3.83 (1.23)	4.16 (1.01)	3.90 (1.62)	4.03 (1.18)
Transfer items	Immediate posttest	4.74 (1.10)	4.88 (1.06)	4.69 (2.25)	5.44 (1.35)
	Delayed posttest	4.27 (1.50)	5.18 (1.18)	5.00 (2.07)	5.21 (1.24)
Mental effort during learning		3.84 (1.10)	4.05 (1.11)	3.97 (1.05)	4.48 (0.85)

Note. Analysis 1 concerns the pretest to immediate posttest analysis for all students present during session 1 and analysis 2 concerns the immediate posttest to delayed posttest analysis for all students present during both sessions.

25% of the open questions of the immediate posttest, blind to student identity and condition. Because the intra-class correlation coefficient was high (.931 for learning test items; .929 for transfer test items), the remainder of the tests was scored by one rater (the third author) and this rater's scores were used in the analyses.

The mean score on the posttest learning items was 62.5% (SD = 19.06) and reliability of these items was .36 on the pretest, .45 on the posttest and .52 on the delayed posttest (Cronbach's alpha). Again, the low reliability on the pretest might be explained by the fact that a lack of prior knowledge requires guessing of answers, resulting in low inter-item correlations and subsequently a low Cronbach's alpha. Moreover, caution is warranted in interpreting these reliabilities because a sample size as in our study does not seem to produce precise alpha coefficients (e.g. Charter, 2003). The mean score on the posttest transfer items was 32.2% (SD = 25.55) and reliability of these items was .36 on the posttest and .30 on the delayed posttest (Cronbach's alpha). In view of this low reliability, which can probably partly be explained by floor effects at both tests for one of our transfer task categories (i.e. Wason selection), we decided not to report the test statistics of the analyses on transfer performance. Descriptive statistics can be found in Tables 5 and 6.

Results

In all analyses reported below, a significance level of .05 was used. Partial eta-squared (η_p^2) is reported as a measure of effect size for the ANOVAs for which 0.01 is considered small, 0.06 medium, and 0.14 large. On our OSF-project page we presented the intention-to-treat (i.e. all participants who entered the study) analyses, which did not reveal noteworthy differences with the compliant-only analyses. As it might have been of influence that half of the students had some prior knowledge as they participated in a study that included similar heuristics-and-biases tasks in the first year of their curriculum, we included the factor Site in all analyses.

Check on condition equivalence and time-on-task

Preliminary analyses confirmed that there were no significant differences between the conditions in educational background, $\chi^2(9) = 10.00$, $p = .350$; gender, $\chi^2(3) = .318$, $p = .957$, or performance on the pretest, time spent on the pretest, and mental effort invested on the pretest items (maximum $F =$ 1.30, maximum $\eta_p^2 = .03$). A one-way ANOVA indicated that there were no significant differences in time-on-task (in seconds) spent on practice of the instruction tasks, $F(3, 116) = 1.73$, $p = .165$, $d = .016$.[7]

[7] The relatively low reliabilities of the learning items should be taken into account.

Table 6. Means (SD) of Test performance per task (max. score 1) per Condition of Experiment 2.

		Instructional conditions			
		Blocked Examples	Blocked Problems	Interleaved Examples	Interleaved Problems
Syllogism 1	Pretest	0.60 (0.51)	0.51 (0.52)	0.60 (0.51)	0.67 (0.48)
	Immediate posttest	0.45 (0.51)	0.51 (0.51)	0.53 (0.52)	0.54 (0.51)
	Delayed posttest	0.53 (0.52)	0.67 (0.48)	0.88 (0.53)	0.75 (0.44)
Syllogism 2	Pretest	0.15 (0.37)	0.40 (0.50)	0.13 (0.35)	0.26 (0.44)
	Immediate posttest	0.70 (0.47)	0.51 (0.51)	0.87 (0.36)	0.56 (0.50)
	Delayed posttest	0.53 (0.52)	0.64 (0.49)	0.75 (0.46)	0.78 (0.42)
Syllogism 3	Pretest	0.35 (0.49)	0.40 (0.50)	0.33 (0.49)	0.31 (0.47)
	Immediate posttest	0.50 (0.51)	0.37 (0.49)	0.67 (0.49)	0.46 (0.51)
	Delayed posttest	0.53 (0.52)	0.55 (0.51)	0.38 (0.52)	0.50 (0.51)
Base-rate 1	Pretest	0.30 (0.47)	0.35 (0.48)	0.33 (0.49)	0.46 (0.51)
	Immediate posttest	0.90 (0.31)	0.53 (0.51)	0.73 (0.46)	0.64 (0.49)
	Delayed posttest	0.87 (0.35)	0.61 (0.50)	0.75 (0.46)	0.75 (0.44)
Base-rate 2	Pretest	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Immediate posttest	0.30 (0.47)	0.05 (0.21)	0.47 (0.52)	0.03 (0.16)
	Delayed posttest	0.20 (0.41)	0.03 (0.17)	0.25 (0.46)	0.00 (0.00)
Base-rate 3	Pretest	0.85 (0.37)	0.63 (0.49)	0.80 (0.41)	0.77 (0.43)
	Immediate posttest	0.95 (0.22)	0.86 (0.35)	0.87 (0.35)	0.95 (0.22)
	Delayed posttest	1.00 (0.00)	0.73 (0.45)	0.75 (0.46)	0.91 (0.30)
Conjunction 1	Pretest	0.20 (0.41)	0.35 (0.48)	0.33 (0.49)	0.41 (0.50)
	Immediate posttest	0.60 (0.50)	0.84 (0.37)	0.87 (0.35)	0.82 (0.39)
	Delayed posttest	1.00 (0.00)	0.73 (0.45)	0.75 (0.91)	0.91 (0.30)
Conjunction 2	Pretest	0.45 (0.50)	0.60 (0.50)	0.60 (0.51)	0.72 (0.46)
	Immediate posttest	0.75 (0.44)	0.81 (0.39)	0.87 (0.35)	0.87 (0.34)
	Delayed posttest	0.80 (0.41)	0.85 (0.36)	0.89 (0.33)	0.91 (0.30)
Conjunction 3	Pretest	0.55 (0.51)	0.77 (0.43)	0.87 (0.35)	0.79 (0.41)
	Immediate posttest	1.00 (0.00)	0.98 (0.15)	1.00 (0.00)	0.95 (0.22)
	Delayed posttest	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.97 (0.18)
Wason selection 1	Immediate posttest	0.07 (0.26)	0.09 (0.29)	0.11 (0.33)	0.13 (0.37)
	Delayed posttest	0.00 (0.00)	0.09 (0.29)	0.22 (0.44)	0.09 (0.30)
Wason selection 2	Immediate posttest	0.13 (0.35)	0.21 (0.42)	0.33 (0.50)	0.31 (0.47)
	Delayed posttest	0.00 (0.00)	0.06 (0.24)	0.00 (0.00)	0.90 (0.30)
Contingency 1	Immediate posttest	0.60 (0.51)	0.67 (0.48)	0.56 (0.53)	0.56 (0.50)
	Delayed posttest	0.80 (0.41)	0.76 (0.44)	0.78 (0.44)	0.69 (0.47)
Contingency 2	Immediate posttest	0.47 (0.52)	0.52 (0.51)	0.56 (0.53)	0.53 (0.51)
	Delayed posttest	0.80 (0.41)	0.88 (0.33)	0.56 (0.53)	0.72 (0.46)

Note: The reported immediate posttest means are based on analysis 1, that is, the pretest to immediate posttest analysis for all students present during session 1.

Performance on learning items

Performance data are presented in Table 5 and 6 and omnibus test statistics in Table 7 (statistics of follow-up analyses are presented in text). The data on learning items were analysed with two 2 × 2 × 2 × 2 mixed ANOVAs with Test Moment (analysis 1: pretest and immediate posttest; analysis 2: immediate posttest and delayed posttest) as within-subjects factor and Practice Schedule (interleaved and blocked), Practice-task Format (worked examples and practice problems), and Site (low prior knowledge and higher prior knowledge learners) as between-subjects factors. In line with Hypothesis 1, the pretest-immediate posttest analysis showed a main effect of Test Moment on learning outcomes: participants performed better on the immediate posttest ($M = 61.40$, $SE = 1.49$) than on the pretest ($M = 46.13$, $SE = 1.59$).

Contrary to Hypothesis 3a (see Table 1 for a schematic overview of the hypotheses), the results did not reveal a significant main effect of Practice Schedule, nor an interaction with Test Moment, indicating that interleaved practice had no differential effect. We did find an interaction effect between Test Moment and Practice-task Format: in line with Hypothesis 5a, there was a higher pretest to immediate posttest performance gain for worked examples ($M_{\text{pre}} = 38.79$; $M_{\text{immediate}} = 71.96$) than for practice problems ($M_{\text{pre}} = 41.71$; $M_{\text{immediate}} = 58.24$), $F(1,\ 109) = 22.18$, $p < .001$, $\eta_p^2 = .17$. In contrast to Hypothesis 6a, the results did not reveal an interaction between Practice Schedule and Practice-task Format, nor an interaction between Practice Schedule, Practice-task Format, and Test Moment.

However, there was a main effect of Site, with higher-knowledge learners performing better ($M = 60.95$, $SE = 2.00$) than low-knowledge learners ($M = 44.39$, $SE = 1.97$). Moreover, we found an interaction between Test Moment and Site, with a higher increase in learning outcomes for low-knowledge learners ($M_{\text{pre}} = 29.36$, $SE = 2.25$; $M_{\text{immediate}} = 59.43$, $SE = 2.31$) compared to higher-knowledge learners

Table 7. Results Mixed ANOVAs experiment 2

		Test performance			Mental effort		
	ANOVA	F-test (df)	p*	η_p^2	F-test (df)	p*	η_p^2
Learning items							
Analysis 1: Pretest – Immediate Posttest	Test Moment	198.07 (1,109)	< .001*	.65	0.55 (1,108)	.459	.01
	Test Moment × Practice Schedule	1.05 (1,109)	.308	.01	0.00 (1,108)	.971	.00
	Test Moment × Practice-task Format	22.18 (1,109)	<.001*	.17	0.81 (1,108)	.370	.02
	Test Moment × Practice Schedule × Practice-task Format	0.35 (1,109)	.558	.00	3.34 (1,108)	.070	.03
	Test Moment × Site	8.73 (1,109)	.004*	.07	2.50 (1,108)	.117	.02
	Test Moment × Site × Practice Schedule	0.30 (1,109)	.584	.00	5.58 (1,108)	.020*	.05
	Test Moment × Site × Practice-task Format	6.04 (1,109)	.016*	.05	1.27 (1,108)	.262	.01
	Test Moment × Site × Practice Schedule × Practice-task Format	0.97 (1,109)	.326	.01	1.37 (1,108)	.244	.01
	Practice Schedule	1.42 (1,109)	.236	.01	0.78 (1,108)	.378	.01
	Practice-task Format	3.70 (1,109)	.057	.03	2.54 (1,108)	.114	.02
	Practice Schedule × Practice-task Format	0.06 (1,109)	.806	.00	1.01 (1,108)	.316	.01
	Site	34.79 (1,109)	<.001*	.24	2.18 (1,108)	.143	.02
	Site × Practice Schedule	2.27 (1,109)	.135	.02	0.03 (1,108)	.855	.00
	Site × Practice-task Format	1.73 (1,109)	.191	.02	0.72 (1,108)	.398	.01
	Site × Practice Schedule × Practice-task Format	1.12 (1,109)	.292	.01	0.63 (1,108)	.430	.01
Analysis 2: Immediate – Delayed Posttest	Test Moment	6.07 (1,80)	.016*	.07	0.65 (1,79)	.422	.01
	Test Moment × Practice Schedule	0.01 (1,80)	.943	.00	0.62 (1,79)	.432	.01
	Test Moment × Practice-task Format	1.29 (1,80)	.260	.02	1.15 (1,79)	.286	.01
	Test Moment × Practice Schedule × Practice-task Format	0.58 (1,80)	.450	.00	7.50 (1,79)	.008*	.09
	Test Moment × Site	0.49 (1,80)	.485	.00	3.13 (1,79)	.081	.04
	Test Moment × Site x Practice Schedule	0.80 (1,80)	.375	.00	0.11 (1,79)	.744	.00
	Test Moment × Site x Practice-task Format	0.02 (1,80)	.898	.01	0.87 (1,79)	.354	.01
	Test Moment × Site x Practice Schedule × Practice-task Format	0.59 (1,80)	.444	.01	0.13 (1,79)	.718	.00
	Practice Schedule	0.00 (1,80)	.984	.00	0.16 (1,79)	.693	.00
	Practice-task Format	1.29 (1,80)	.260	.02	1.27 (1,79)	.264	.02
	Practice Schedule × Practice-task Format	1.50 (1,80)	.225	.02	0.24 (1,79)	.623	.00
	Site	12.72 (1,80)	.001*	.14	0.17 (1,79)	.686	.00
	Site × Practice Schedule	0.19 (1,80)	.891	.00	0.01 (1,79)	.909	.00
	Site × Practice-task Format	0.07 (1,80)	.800	.00	0.02 (1,79)	.878	.00
	Site × Practice Schedule × Practice-task Format	7.01 (1,80)	.010*	.08	0.14 (1,79)	.715	.00
Practice tasks	Practice Schedule	–	–	–	1.34 (1,109)	.250	.01
	Practice-task Format	–	–	–	2.34 (1,109)	.129	.02
	Practice Schedule × Practice-task Format	–	–	–	0.69 (1,109)	.409	.01
	Site	–	–	–	1.11 (1,109)	.294	.01
	Site × Practice Schedule	–	–	–	0.15 (1,109)	.698	.00
	Site × Practice-task Format	–	–	–	0.32 (1,109)	.572	.00
	Site × Practice Schedule × Practice-task Format	–	–	–	0.62 (1,109)	.431	.01

Note: Analysis 1 concerns the pretest to immediate posttest analysis for all students present during session 1 and analysis 2 concerns the immediate posttest to delayed posttest analysis for all students present at both sessions.

*$p < .05$.

($M_{pre} = 51.14$, $SE = 2.38$; $M_{immediate} = 70.77$, $SE = 2.34$). Interestingly, our results revealed an interaction between Test Moment, Practice-task Format, and Site. Follow-up analyses revealed that low-knowledge learners showed a larger increase in learning outcomes when they practiced with worked examples ($M_{pre} = 27.58$, $SE = 2.83$; $M_{immediate} = 70.30$, $SE = 4.28$) compared to practice problems ($M_{pre} = 31.14$, $SE = 2.63$; $M_{immediate} = 48.55$, $SE = 2.94$), $F(1, 53) = 22.17$, $p < .001$, $\eta_p^2 = .30$. For higher-knowledge learners, the differences in learning gains between the worked examples and practice problems conditions were no longer significant, $F(1, 56) = 3.00$, $p = .089$, $\eta_p^2 = .05$.

The second analysis – to test whether our results are still present after two weeks – showed a significant main effect of Test Moment: participants' performance on learning items improved from immediate ($M = 63.13$, $SE = 2.19$) to delayed ($M = 67.71$, $SE = 2.31$) posttest. In contrast to Hypotheses 3a, 5a, and 6a respectively, there was no main effect of Practice Schedule, no main effect of Practice-task Format, no interaction between Practice Schedule and Practice-task Format, nor interactions with Test Moment. Again, there was a main effect of Site: higher-knowledge learners performed higher on learning items ($M = 72.73$, $SE = 2.49$) than low-knowledge learners ($M = 58.11$, $SE = 3.26$). Furthermore, an interaction between Practice Schedule, Practice-task Format, and Site was found. Follow-up analyses revealed that, for low-knowledge learners practice in a blocked schedule worked best with worked examples compared to practice problems ($M_{WE} = 69.14$, $SE = 5.78$; $M_{PS} = 47.57$, $SE = 4.34$), while in an interleaved schedule practice problems were more beneficial ($M_{WE} = 52.78$, $SE = 12.27$; $M_{PS} = 62.96$, $SE = 5.01$), $F(1, 35) = 4.43$, $p = .043$, $\eta_p^2 = .11$. There was no significant interaction between Practice Schedule and Practice-task Format for higher-knowledge learners, $F(1, 45) = 1.87$, $p = .178$, $\eta_p^2 = .04$. No other interaction effects were found.

Mental effort during learning

Mental effort data are presented in Table 5 and omnibus test statistics in Table 7. Contrary to Hypotheses 2 and 4, respectively, a 2 (Practice Schedule: interleaved and blocked) × 2 (Practice-task Format: worked examples and practice problems) × 2 (Site: low prior knowledge learners and higher prior knowledge learners) factorial ANOVA on the mental effort during practice data revealed no main effects of Practice Schedule and Practice-task Format, nor an interaction between Practice Schedule and Practice-task Format was found. Moreover, no main effect of Site, nor interactions between Practice Schedule, Practice-task Format, and Site were found.

Mental effort during test

Our pretest-immediate posttest analyses on effort invested on *learning items* showed no main effects of Practice Schedule (Question 7a) and Practice-task Format (Question 8a), nor an interaction between Practice Schedule and Practice-task Format. The results did reveal a significant interaction between Test Moment, Practice Schedule, and Site, but follow-up analyses revealed no significant interactions between Test Moment and Practice Schedule for both sites (maximum $F = 3.47$, maximum $\eta_p^2 = .06$). No main effects of Test Moment and Site, nor other significant interactions were found.

Our second analysis – to test whether our results were still present after two weeks – showed no main effects of Practice Schedule (Question 7b) and Practice-task Format (Question 8b), nor an interaction between Practice Schedule and Practice-task Format. However, a three-way interaction between Test Moment, Practice Schedule, and Practice-task Format was found. Follow-up analyses revealed that interleaved practice with worked examples resulted in an immediate posttest – delayed posttest increase in effort investment ($M_{immediate} = 3.58$; $M_{delayed} = 3.97$) and with practice problems in an immediate posttest – delayed posttest decrease in effort investment ($M_{immediate} = 4.45$; $M_{delayed} = 4.07$), $F(1, 36) = 4.21$, $p = .047$, $\eta_p^2 = .11$. There was no significant difference in immediate posttest – delayed posttest effort investment between the practice-task format conditions when practiced in a blocked schedule, $F(1, 43) = 2.74$, $p = .105$, $\eta_p^2 = .06$. No main effects of Test Moment and Site, nor other interactions were found.

Our analyses on effort invested in *transfer items* revealed no main effects of Practice Schedule, Practice-task Format, Test Moment, or Site. Moreover, there were no significant interaction effects.

Interim summary

The results of Experiment 2 provide converging evidence with Experiment 1. Again, we did not find any indications that interleaved practice would be more beneficial than blocked practice for learning, either in itself or as a function of task format. There was again a benefit of studying worked examples over solving problems, but – as was to be expected – this was limited to participants who had low prior knowledge (i.e. had not participated in a study that included similar heuristics and biases tasks in the first year of their curriculum).

General discussion

Previous research has demonstrated that providing students with explicit CT-instructions combined with practice on domain-relevant tasks is beneficial for learning to reason in an unbiased manner (e.g. Heijltjes et al., 2015) but not for transfer to new tasks. Therefore, the present experiments investigated whether creating contextual interference in instruction through interleaved practice – which has been proven effective in other and similar domains – would promote both learning and transfer of reasoning skills.

In line with our expectations and consistent with earlier research (e.g. Van Peppen et al., 2018; Heijltjes et al., 2015), both experiments support the finding that explicit instructions combined with practice improves learning of unbiased reasoning (Hypothesis 1), as we found pretest to immediate posttest gains on practiced tasks in all conditions, which remained stable on the delayed posttest after two weeks. This is in line with the idea of Stanovich (2011) that providing students with relevant mindware (i.e. knowledge bases, rules, procedures and strategies; Perkins, 1995) and stimulating them to inhibit incorrectly used intuitive responses (i.e. Type 1 processing, e.g. Evans, 2008; Kahneman & Klein, 2009; Stanovich, 2011; Stanovich et al., 2016) and to replace these with more analytical and effortful reasoning (i.e. Type 2 processing) is useful to prevent biases in reasoning and decision-making. However, the scores were not particularly high (i.e. up to 73% accuracy), so there is still room for improvement. The performance gain on practiced tasks suggests that having learners *repeatedly* retrieve to-be-learned material (i.e. repeated retrieval practice: e.g. Karpicke & Roediger, 2007) may be a promising method to further enhance learning to avoid biased reasoning.

Contrary to our hypotheses, we did not find any indications that interleaved practice would improve learning more than blocked practice (Hypothesis 3a), regardless of whether they practiced with worked examples or problem-solving tasks (Hypothesis 6a). These findings are in contrast to previous studies that demonstrated that interleaved practice is effective for establishing both learning and transfer in other domains and with other complex judgment tasks (e.g. Likourezos et al., 2019). Moreover, they are contrary to the finding of Paas and Van Merriënboer (1994) that high variability during practice with geometrical problems produced test performance benefits when students studied worked examples, but not when they solved practice problems. Unfortunately, we were not able to test our hypotheses regarding transfer performance (Hypothesis 3b/6b). Therefore, it is unknown whether interleaved practice – either in itself or as a function of task-format – would be beneficial for transfer of unbiased reasoning. However, given that the transfer scores were overall rather low, we can assume the overall effect of instruction and practice (if present at all) would seem to be limited.

One of the more interesting findings to emerge from this study, however, is that the *worked example effect* (e.g. Paas & Van Gog, 2006; Renkl, 2014) also applies to CT-tasks. Moreover, this was found even though the instructions that preceded the practice tasks already included two worked examples. As most of the studies on the worked example effects used pure practice conditions or gave minimal instructions prior to practice, these examples could have helped students in the problem-solving conditions perform better on the practice problems; nevertheless, we still found a worked example effect. To the best of our knowledge, the results of Experiment 1 demonstrated for the first time in CT-instruction a benefit of studying worked examples over solving problems on learning outcomes, reached with less effort during the tests (i.e. more effective and efficient, Van Gog & Paas, 2008). Experiment 2 replicated the worked example effect (i.e. more effective than solving problems) and demonstrated that this was the case for novices, but not for learners with relatively more prior knowledge. This observation supports findings regarding the *expertise reversal effect* (e.g. Kalyuga, 2007; Kalyuga et al., 2003, 2012), which shows that while instructional strategies that assist learners in developing cognitive schemata are effective for low-knowledge learners, they are often not effective (or may even be detrimental) for higher-knowledge learners. As far as we know, our second experiment was the first to actually vary both level of guidance (i.e. practice-task format) and level of expertise along with practice schedule and, thus, our study provides a first step in exploring the interactions between these factors. However, caution is warranted in interpreting this finding since our sample size was relatively small. It would be interesting in future research to manipulate students' level of expertise to actually demonstrate a causal relationship between expertise and the effect of studying worked examples on learning outcomes in CT-instruction.

Admittedly, our explanation for the worked example effect would have been more compelling if we had included measures of the separate types of cognitive load (although it seems very challenging to distinguish between different types of cognitive load and available instruments, e.g. the rating scale developed by Leppink et al., 2013, would be too long to apply after each task). There are theoretical reasons to assume that the amount of strategy information given in the worked examples resulted in lower extraneous load and higher germane load compared to solving problems, but in that case, the total load experienced by the learners (as reflected in their invested mental effort) may not differ between conditions. Moreover, studies in which worked examples were compared to practice problems with feedback consisting of or resembling the supportive features of worked examples, still showed a worked example effect. Paas and Van Merriënboer (1994), for instance, showed that training with worked examples was more beneficial for learning than training with practice-problems that were followed by correct-answer feedback and worked examples. Moreover, training with worked examples required less time and was perceived as less effortful. In line with these findings, both McLaren et al. (2016) and Schwonke et al. (2009) demonstrated that worked examples were less time consuming without a loss or even a gain in learning outcomes compared to tutored learning by problem-solving (i.e. clear efficiency benefits of worked example study). It is important to note that the worked example effect does apply to learners who have little prior knowledge while it disappears for learners with high prior knowledge (cf. expertise reversal effect). In the current study, learners were provided with prior knowledge during the initial CT-instructions. We nevertheless revealed a worked example effect. Hence, it seems that participants did not develop such expertise that the positive effect of worked examples disappeared. This allowed learners to still take advantage from the information provided in the worked examples. If learners have not learned from the initial instructions at all (which is unlikely given previous research), the elaborate information in the worked examples may have helped them to apprehend the needed approach to problem solving, while the information in the practice problems and subsequent feedback whether the answer was correct or incorrect could at best hint at what might be a reasonable approach.

Finally, one could argue that the unequal cell distribution (i.e. higher exclusion in worked examples conditions compared to practice problems conditions based on reading time of instructions) may indicate that students' motivation may have been the basis for the worked example effect. However, our intention-to-treat analyses still revealed a worked example effect and, therefore, this possible explanation does not seem convincing. Yet, this points to another remarkable finding, that is, that worked examples were more beneficial for learning than problems, even if the examples were minimally read; possibly, students quickly located and processed the relevant information in the examples.

Although we have to speculate, a possible explanation for the absence of an interleaved practice effect on learning outcomes might lie in the distinctiveness between the task categories, which may have been greater than in previous studies. Effects of interleaved practice only occur if task categories differ and require different problem-solving procedures. However, as reflection on the to-be-used procedures is what causes the beneficial effect of interleaved practice (e.g. Barreiros et al., 2007; Rau et al., 2010), distinctiveness between categories should not be too high because learners then immediately recognise what procedure to apply. It seems possible that the task categories used in the present study were the same at a high level but that the mindware needed for each category differed too much. If so, determining the nature of each task was relatively easy and intertask comparing was not necessary. It should be noted, though, that this was not expected in advance and that arguing that the distinctiveness between task categories was too high *after* we know the results is risky, because of hindsight bias (Fischhoff, 1975).

Another, again speculative, possible explanation for the absence of an interleaving effect on learning outcomes, might be that the surface characteristics within the practice-task categories were so different (especially for the base-rate items) that students in the blocked practice condition did not realise that strategies could be reused in subsequent tasks of that category. This suggestion is supported by the performance differences on base-rate and syllogistic reasoning items, although the latter is more likely due to differences in difficulty. To reiterate, in the blocked practice condition, participants practiced with three tasks of one category at a time before the next (e.g. AAABBBCCC), whereas in the interleaved practice condition, participants practiced with the nine heuristics-and-biases tasks in a mixed sequence (e.g. ABCBACBCA). As such, students in the blocked practice condition might have been stimulated as much as students in the interleaved practice condition to stop and think about new problem-solving strategies, especially in base-rate tasks. It seems possible that interleaved practice is useful for practice within a task category in which surface characteristics are similar to each other and problem-solving procedures differ slightly (e.g. syllogistic reasoning tasks), but further research should be undertaken to investigate this.

Additionally, a recent meta-analysis (Brunmair & Richter, 2019) has shown that the strength of interleaved practice effects varies widely between types of learning materials. Interleaved practice seems to work well in inductive learning, when the stimuli are complex, when categories are difficult to discriminate, and when the similarity of exemplars within categories is low. Given the pervasiveness of induction, it is surprising that educationally relevant materials are clearly underrepresented in the interleaved practice research. The present study was the first to address interleaved practice effects with heuristics-and-biases tasks, and seems to indicate that interleaved practice is not beneficial for learning of this type of task. Hence, it would be interesting for future studies to investigate the generalizability of the interleaved practice effects and whether it is restricted to specific types and combinations of learning materials. More generally, our findings raise questions about the preconditions of instructional strategies that are known to foster generative processing (e.g. desirable difficulties; Bjork, 1994). Instructional strategies, such as interleaved practice, depend highly on the implementation, the measure of learning outcomes, and the specific characteristics of the learning materials. Further research on the exact boundary conditions is therefore recommended to accurately inform educational practice.

Moreover, it should be noted that the relatively low reliabilities, implying high amounts of measurement error, of our learning test items might have played a crucial role as it largely decreased the power to detect intervention effects (Cleary et al., 1970; Kanyongo et al., 2007; Schmidt & Hunter, 1996). Although sample sizes as in studies like this do not seem to produce sufficiently precise alpha coefficients (e.g. Charter, 2003), the possibility that the items were not sufficiently related or that students do not see the overlap between the items should be taken into account. In this study, the low levels of reliability can probably be explained in terms of multidimensionality of the tests encompassing several heuristics-and-biases tasks, a factor often ignored in current research. Performance on these tasks depends not only on the extent to which that task elicits a bias (resulting from heuristic reasoning), but also on the extent to which one possesses the requisite mindware (e.g. rules or logic or probability). Thus, systematic variance in performance on such tasks can either be explained by a person’s use of heuristics or his/her available mindware. If it differs per item to what extent a correct answer depends on these two aspects, there may not be a common factor explaining all interrelationships between the measured items. Future research, therefore, would need to find ways to improve CT measures (i.e. decrease random measurement error) or should utilise measures known to have acceptable levels of reliability (LeBel & Paunonen, 2011). The latter option seems challenging, however, as multiple studies report rather low levels of reliability of tests consisting of heuristics and biases tasks (Aczel et al., 2015; Bruine de Bruin et al., 2007; West et al., 2008) and revealed concerns with the reliability of widely used standardised CT tests, particularly with regard to subscales (Bernard et al., 2008; Ku, 2009; Leppa, 1997; Liu et al., 2014).

One could argue to what extent the tests accurately assessed the more general cognitive capacity “avoiding bias in reasoning” (i.e. unbiased reasoning). Bias refers to systematic deviations from a norm when choosing actions or estimating probabilities (Stanovich et al., 2016; Tversky & Kahneman, 1974). In the current study, unbiased reasoning was operationalised as performance on classical heuristics-and-biases tasks, in which an intuitively cued heuristic response conflicts normative models of CT as set by formal logic and probability theory. Heuristics-and-biases tasks have been used for decades to measure unbiased reasoning (e.g.

Baron, 2008; Evans, 2003; Gigerenzer & Hug, 1992; Heijltjes et al., 2014a; Heijltjes et al., 2014b, 2015; Tversky & Kahneman, 1974; Stanovich et al., 2016; Stanovich & West, 2000; Tversky & Kahneman, 1983; Van Brussel et al., 2020; Wasserman et al., 1990; West et al., 2008). Several studies demonstrated associations between people's performance on heuristics-and-biases tasks and how they reason in more realistic settings (e.g. medical decision making: Arkes, 2013) and other real-world correlates (e.g. risk behaviours: Toplak et al., 2017). Hence, participants' performance on these heuristics-and-biases tasks presumably offers a realistic view of everyday reasoning (see for example, Gilovich et al., 2002). Relevant next steps would be to investigate how bias in reasoning can be prevented in daily settings and what the effects of instruction/practice are on other aspects of CT.

To conclude, the present experiments provide evidence that worked examples can be effective for novices' learning to avoid biased reasoning. However, there were no indications that practice in an interleaved schedule – with worked examples or practice problems – enhances performance on heuristics-and-biases tasks. These findings suggest that the nature or the combination of the task categories may be a boundary condition for effects of interleaved practice on learning and transfer. Further research should be undertaken to investigate what the exact boundary conditions of effects of interleaved practice are and to provide more insight into the expertise-reversal effect in CT-instruction. Moreover, future research could investigate whether other types of (generative) activities would be beneficial for establishing learning and transfer of unbiased reasoning and whether it is feasible at all to teach students to inhibit Type 1 processing and to recognise when Type 2 processing is needed. It is important to continue the search for effective methods to foster transfer, because biased reasoning can have huge negative consequences in situations in both daily life and complex professional environments.

Acknowledgments

This work was supported by the Netherlands Organisation for Scientific Research under Grant 409-15-203.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Netherlands Organisation for Scientific Research under [grant number 409-15-203]; Nederlandse Organisatie voor Wetenschappelijk Onderzoek

Data deposition

The datasets and script files are stored on an Open Science Framework (OSF) page for this project, see osf.io/a9czu.

ORCID

Peter P. J. L. Verkoeijen http://orcid.org/0000-0002-8085-5038

References

Abel, M., & Roediger, H. L. (2017). Comparing the testing effect under blocked and mixed practice: The mnemonic benefits of retrieval practice are not affected by practice format. *Memory & Cognition*, *45*(1), 81–92. https://doi.org/10.3758/s13421-016-0641-8

Aczel, B., Bago, B., Szollosi, A., Foldes, A., & Lukacs, B. (2015). Measuring individual differences in decision biases: Methodological considerations. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01770

Ajayi, T., & Okudo, J. (2016). Cardiac arrest and gastrointestinal bleeding: A case of medical heuristics. *Case Reports in Medicine*, *2016*. https://doi.org/10.1155/2016/9621390

Albaret, J. M., & Thon, B. (1998). Differential effects of task complexity on contextual interference in a drawing task. *Acta Psychologica*, *100*(1-2), 9–24. https://doi.org/10.1016/S0001-6918(98)00022-5

Arkes, H. R. (2013). The consequences of the hindsight bias in medical decision making. *Current Directions in Psychological Science*, *22*(5), 356–360. https://doi.org/10.1177/0963721413489988

Baron, J. (2008). *Thinking and deciding* (4th ed.). Cambridge University Press.

Barreiros, J., Figueiredo, T., & Godinho, M. (2007). The contextual interference effect in applied settings. *European Physical Education Review*, *13*(2), 195–208. https://doi.org/10.1177/1356336X07076876

Battig, W. F. (1978). The flexibility of human memory. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing and human memory* (pp. 23–44). Erlbaum.

Bernard, R. M., Zhang, D., Abrami, P. C., Sicoly, F., Borokhovski, E., & Surkes, M. A. (2008). Exploring the structure of the Watson–Glaser critical thinking appraisal: One scale or many subscales? *Thinking Skills and Creativity*, *3*(1), 15–22. https://doi.org/10.1016/j.tsc.2007.11.001

Billings, L., & Roberts, T. (2014). *Teaching critical thinking: Using seminars for 21st century literacy*. Routledge.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology*, *92* (5), 938–956. https://doi.org/10.1037/0022-3514.92.5.938

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029–1052. https://doi.org/10.1037/bul0000209

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition*, *41*(5), 671–682. https://doi.org/10.3758/s13421-012-0291-4

Carvalho, P. F., & Goldstone, R. L. (2019). When does interleaving practice improve learning. In J. Dunlosky, & K. Rawson (Eds.), *The Cambridge Handbook of Cognition and education* (pp. 183–208). Cambridge University Press.

Charter, R. A. (2003). Study samples are too small to produce sufficiently precise reliability coefficients. *The Journal of General Psychology*, *130*(2), 117–129. https://doi.org/10.1080/00221300309601280

Cleary, T. A., Linn, R. L., & Walster, G. W. (1970). Effect of reliability and validity on power of statistical tests. *Sociological Methodology*, *2*, 130–138. https://doi.org/10.2307/270786

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd. ed., reprint). Psychology Press.

Darling-Hammond, L. (2010). Teacher education and the American future. *Journal of Teacher Education*, *61*(1-2), 35–47. https://doi.org/10.1177/0022487109348024

De Croock, M. B., & van Merriënboer, J. J. (2007). Paradoxical effects of information presentation formats and contextual interference on transfer of a complex cognitive skill. *Computers in Human Behavior*, *23*(4), 1740–1761. https://doi.org/10.1016/j.chb.2005.10.003

De Croock, M. B., van Merriënboer, J. J., & Paas, F. G. (1998). High versus low contextual interference in simulation-based training of troubleshooting skills: Effects on transfer performance and invested mental effort. *Computers in Human Behavior*, *14*(2), 249–267. https://doi.org/10.1016/S0747-5632(98)00005-3

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

Elia, F., Apra, F., Verhovez, A., & Crupi, V. (2016). "First, know thyself": Cognition and error in medicine. *Acta Diabetologica*, *53*(2), 169–175. https://doi.org/10.1007/s00592-015-0762-8

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. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

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*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fischhoff, B. (1975). Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 288–299. https://doi.org/10.1037/0096-1523.1.3.288

Flores, K. L., Matkin, G. S., Burbach, M. E., Quinn, C. E., & Harding, H. (2012). Deficient critical thinking skills among college graduates: Implications for leadership. *Educational Philosophy and Theory*, *44*(2), 212–230. https://doi.org/10.1111/j.1469-5812.2010.00672.x

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition*, *43*(2), 127–171. https://doi.org/10.1016/0010-0277(92)90060-U

Gilovich, T, Griffin, D, & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist*, *53*(4), 449–455. https://doi.org/10.1037/0003-066X.53.4.449

Heijltjes, A., Van Gog, T., Leppink, J., & Paas, F. (2014a). 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. (2014b). 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

Helsdingen, A., Van Gog, T., & Van Merriënboer, J. (2011a). The effects of practice schedule and critical thinking prompts on learning and transfer of a complex judgment task. *Journal of Educational Psychology*, *103*(2), 383. https://doi.org/10.1037/a0022370

Helsdingen, A. S., Van Gog, T., & van Merriënboer, J. J. (2011b). The effects of practice schedule on learning a complex judgment task. *Learning and Instruction*, *21*(1), 126–136. https://doi.org/10.1016/j.learninstruc.2009.12.001

Hoffman, B., & Schraw, G. (2010). Conceptions of efficiency: Applications in learning and problem solving. *Educational Psychologist*, *45*(1), 1–14. https://doi.org/10.1080/00461520903213618

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526. https://doi.org/10.1037/a0016755

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. https://doi.org/10.1037/h0034747

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, *19*(4), 509–539. https://doi.org/10.1007/s10648-007-9054-3

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

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, *38*(1), 23–31. https://doi.org/10.1207/S15326985EP3801_4

Kalyuga, S., Rikers, R., & Paas, F. (2012). Educational implications of expertise reversal effects in learning and performance of complex cognitive and sensorimotor skills. *Educational Psychology Review*, *24*(2), 313–337. https://doi.org/10.1007/s10648-012-9195-x

Kanyongo, G. Y., Brook, G. P., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods*, *6*(1), 81–90. https://doi.org/10.22237/jmasm/1177992480

Karpicke, J. D., & Roediger III, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*, 151–162. https://doi.org/10.1016/j.jml.2006.09.004

Koehler, D. J., Brenner, L., & Griffin, D. (2002). The calibration of expert judgment: Heuristics and biases beyond the labratory. In T. Gilovich, D. W. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 686–715). Cambridge University Press.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Ku, K. Y. L. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity*, *4*(1), 70–76. https://doi.org/10.1016/j.tsc.2009.02.001

Kuhn, D. (2005). *Education for thinking*. Harvard University Press.

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin*, *37*(4), 570–583. https://doi.org/10.1177/0146167211400619

Leppa, C. J. (1997). Standardized measures of critical thinking: Experience with the california critical thinking tests. *Nurse Educator*, *22*(5), 29–33. https://doi.org/10.1097/00006223-199709000-00012

Leppink, J., Paas, F., Van der Vleuten, C. P., Van Gog, T., & Van Merriënboer, J. J. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, *45*(4), 1058–1072. https://doi.org/10.3758/s13428-013-0334-1

Likourezos, V., Kalyuga, S., & Sweller, J. (2019). The variability effect: When instructional variability is advantageous. *Educational Psychology Review*, 1–19. https://doi.org/10.1007/s10648-019-09462-8

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series*, *2014*, 1–23. https://doi.org/10.1002/ets2.12009

Mamede, S., van Gog, T., Van Den Berge, K., Rikers, R. M., Van Saase, J. L., Van Guldener, C., & Schmidt, H. G. (2010). Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *JAMA*, *304*(11), 1198–1203. https://doi.org/10.1001/jama.2010.1276

McLaren, B. M., Van Gog, T., Ganoe, C., Karabinos, M., & Yaron, D. (2016). The efficiency of worked examples compared to erroneous examples, tutored problem solving, and problem solving in computer-based learning environments. *Computers in Human Behavior*, *55*, 87–99. https://doi.org/10.1016/j.chb.2015.08.038

Moxley, S. E. (1979). Schema: The variability of practice hypothesis. *Journal of Motor Behavior*, *11*(1), 65–70. https://doi.org/10.1080/00222895.1979.10735173

Paas, F. (1992). Training strategies for attaining transfer or problem solving skills in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1–4. https://doi.org/10.1207/S15326985EP3801_1

Paas, F., & Van Gog, T. (2006). Optimising worked example instruction: Different ways to increase germane cognitive load. *Learning and Instruction*, *16*(2), 87–91. https://doi.org/10.1016/j.learninstruc.2006.02.004

Paas, F. G., & Van Merriënboer, J. J. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, *86*(1), 122–144. https://doi.org/10.1037/0022-0663.86.1.122.

Perkins, D. (1995). *Outsmarting IQ: The emerging science of learnable intelligence*. Free Press.

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.

Rau, M. A., Aleven, V., & Rummel, N. (2010). Blocked versus interleaved practice with multiple representations in an intelligent tutoring system for fractions. In V. Aleven, &

K. J. Mostow (Eds.), *International conference on intelligent tutoring systems* (pp. 413–422). Springer.

Rau, M. A., Aleven, V., & Rummel, N. (2013). Interleaved practice in multi-dimensional learning tasks: Which dimension should we interleave? *Learning and Instruction*, *23*, 98–114. https://doi.org/10.1016/j.learninstruc.2012.07.003

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

Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). Linking cognitive science to education: Generation and interleaving effects. In *Proceedings of the twenty-seventh annual conference of the cognitive science society* (pp. 1850–1855). Erlbaum.

Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. *Psychonomic Bulletin & Review*, *21*(5), 1323–1330. https://doi.org/10.3758/s13423-014-0588-3

Rohrer, D., Dedrick, R. F., Hartwig, M. K., & Cheung, C.-N. (2019). A randomized controlled trial of interleaved mathematics practice. *Journal of Educational Psychology*, *112*(1), 40–52. https://doi.org/10.1037/edu0000367

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology*, *107*(3), 900–908. https://doi.org/10.1037/edu0000001

Sana, F., Yan, V. X., Kim, J. A., Bjork, E. L., & Bjork, R. A. (2018). Does working memory capacity moderate the interleaving benefit? *Journal of Applied Research in Memory and Cognition*, *7*(3), 361–369. https://doi.org/10.1016/j.jarmac.2018.05.005

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 27 research scenarios. *Psychological Methods*, *1*(2), 199–223. https://doi.org/10.1037/1082-989X.1.2.199

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*(2), 419–440. https://doi.org/10.1006/jmla.2001.2813

Schneider, V. I., Healy, A. F., Ericsson, K. A., & Bourne Jr, L. E. (1995). The effects of contextual interference on the acquisition and retention of logical rules. In A. F. Healy, & L. E. Bourne, Jr (Eds.), *Learning and memory of knowledge and skills: Durability and specificity* (pp. 95–131). Sage Publications, Inc.

Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., & Salden, R. (2009). The worked-example effect: Not an artefact of lousy control conditions. *Computers in Human Behavior*, *25*(2), 258–266. https://doi.org/10.1016/j.chb.2008.12.011

Stanovich, K. E. (2011). *Rationality and the reflective mind.* Oxford University Press.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435

Stanovich, K. E., West, R. K., & Toplak, M. E. (2016). *The rationality quotient: Toward a test of rational thinking*. MIT Press.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Measuring cognitive load. In *Cognitive load theory. Explorations in the learning sciences, instructional systems and performance technologies* (pp. 71–85). Springer.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2017). Real-world correlates of performance on heuristics and biases tasks in a community sample. *Journal of Behavioral Decision Making*, *30*(2), 541–554. https://doi.org/10.1002/bdm.1973

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*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Van Brussel, S., Timmermans, M., Verkoeijen, 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*, Article 101844. https://doi.org/10.1016/j.cedpsych.2020.101844

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. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review*, *22*(2), 155–174. https://doi.org/10.1007/s10648-010-9134-7

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 Merriënboer, J. J. G., De Croock, M. B., & Jelsma, O. (1997). The transfer paradox: Effects of contextual interference on retention and transfer performance of a complex cognitive skill. *Perceptual and Motor Skills*, *84*(3), 784–786. https://doi.org/10.2466/pms.1997.84.3.784

Van Merriënboer, J. J. G., Schuurman, J. G., de Croock, M. B., & Paas, F. (2002). Redirecting learners' attention during training: Effects on cognitive load, transfer test performance and training efficiency. *Learning and*

Instruction, *12*(1), 11–37. https://doi.org/10.1016/S0959-4752(01)00020-2

Van Peppen, L. M., Verkoeijen, P. P. J. L., Heijltjes, A. E., 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

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition*, *39*(5), 750–763. https://doi.org/10.3758/s13421-010-0063-y

Wasserman, E. A., Dorner, W. W., & Kao, S. F. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 509–521. https://doi.org/10.1037/0278-7393.16.3.509

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

Zulkiply, N., & Burt, J. S. (2013). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition*, *41*(1), 16–27. https://doi.org/10.3758/s13421-012-0238-9