How Working Memory Capacity and Shifting Matter for Learning with Worked Examples – A Replication Study

Sarah Bichler[1], Matthias Schwaighofer[1], Matthias Stadler[1], Markus Bühner[1], Samuel Greiff[2], & Frank Fischer[1]

[1]Ludwig-Maximilians Universität München

[2]Université du Luxembourg

Author Note

Sarah Bichler, Matthias Stadler, Frank Fischer, and Markus Bühner, Department of Psychology, Ludwig-Maximilians-Universität München. Samuel Greiff, Institute of Cognitive Science and Assessment, Université du Luxembourg.

This manuscript is based on data used in the doctoral dissertation of the first author. This research has been presented at conferences prior to manuscript submission. Matthias Schwaighofer, who started this research, died on June 10th, 2017.

This research was supported by the Elite Network of Bavaria (project number: K-GS-2012-209).

Correspondence concerning this article should be addressed to Sarah Bichler, Department of Psychology, Ludwig-Maximilians-Universität München, Leopoldstr. 13, 80802 Munich, Germany. E-mail: Sarah.Bichler@psy.lmu.de

Abstract

A previous study found that task shifting and fluid intelligence, but not working memory capacity (WMC) and prior knowledge, influenced the worked example effect (Schwaighofer, Bühner, & Fischer, 2016). To increase confidence in these findings, we report a pre-registered extended replication study of Schwaighofer et al.'s investigation. University students (N = 231, M_{age} = 22.40 (SD = 4.33), 87% women) solved statistical problems with textbook materials presented on a laptop in one of four conditions in a 2 x 2 factorial between-subjects design. We compared worked examples vs. problem-solving (replication) and with vs. without time pressure (extension). Time pressure was added to test whether learners in the original study were able to offload WMC demands, which would explain why the WMC moderation was not found. Results showed that the advantage of worked examples over problem-solving decreased with increasing prior knowledge, suggesting that problem-solving becomes eventually more effective than worked example study. Similarly, the benefit of worked examples over problem-solving decreased with increasing shifting ability of a learner. However, contingencies on WMC or fluid intelligence were not detected. Our extension analysis indicated the worked example effect was also not contingent on WMC even when learners were under time pressure. These findings underline the important role that task shifting might play in scaffolded learning environments and suggest that trading in the focus on WMC for a broader perspective on cognitive architecture provides novel explanations for instructional effectiveness. Our study further highlights the importance of more customized instructional support.

Key words: worked examples, problem-solving, working memory capacity, shifting, fluid intelligence

Educational Impact and Implications Statement

In our study we investigated whether learners' cognitive characteristics influence the effectiveness of instruction. We compared worked example study to problem-solving and found that while worked examples facilitated learning for those learners with lower prior knowledge, learners with higher prior knowledge can handle problem-solving demands. Likewise, when learners are good in shifting their attention between different tasks, a cognitive ability referred to as "task shifting", problem-solving is an aptly suited approach to foster learning but when this ability is low, providing worked examples seems to be vital. These findings underpin the importance of more customized approaches to ultimately achieve effective instruction for all learners.

How Working Memory Capacity and Shifting Matter for Learning with Worked Examples – A Replication Study

By building upon what is known about human cognition, instructional theory has matured over the last decades and developed principles for effective instruction (Ginns & Leppink, 2019; Sweller, 2010; Sweller, van Merriënboer, & Paas, 2019). In particular, fostering complex learning effectively requires instructional design to consider the limited capacity of our working memory (Sweller, 1988; Sweller et al., 2019). Besides working memory capacity (WMC), there are other characteristics of our cognitive architecture that are relevant for learning but have not yet been systematically considered in instructional theory. For instance, it is well established that fluid intelligence predicts academic achievement (e.g., Primi, Ferrão, & Almeida, 2010; Roth et al., 2015) and further, our ability to flexibly switch between executing different tasks, referred to as "shifting" (Miyake et al., 2000), has been shown to positively predict math and reading performance (Yeniad, Malda, Mesman, van IJzendoorn, & Pieper, 2013). We suggest that broadening the view on cognitive architecture may enrich instructional theory as it may provide more detailed insight into cognition and learning which will result in refinement of known or discovery of new instructional principles and may ultimately lead to more effective instruction.

We ground our suggestion in early work on aptitude-treatment interactions that identified fluid intelligence as a moderator of instructional treatment (Snow & Lohman, 1984), as well as in a previous investigation on statistical problem-solving with worked examples finding that the benefit of worked examples over problem-solving was contingent on shifting ability and fluid intelligence, but not detecting such contingencies for WMC (Schwaighofer, Bühner, & Fischer, 2016). These results give reason to contemplate advancing instructional theory from focus on a single cognitive function (WMC) to consideration of multiple cognitive characteristics (e.g., WMC, shifting, reasoning).

Moreover, they suggest task shifting as a crucial component of cognitive architecture within instructionally supported learning, which aligns with most recent suggestions in advancing instructional theory (Sepp, Howard, Tindall-Ford, Agostinho, & Paas, 2019).

We report a pre-registered (https://osf.io/dx6qv/) (Bichler, Bühner, Fischer, Stadler, & Greiff, 2019, October 5) replication study of Schwaighofer and colleagues' (2016) investigation. Although replication is seen as a major pillar of empirical research (Schmidt, 2009), replication studies in psychological research are scarce at best (Makel, Plucker, & Hegarty, 2012). We aim to strengthen Schwaighofer and colleagues' contribution by replicating the experimental conditions of their study (worked examples vs. problem-solving) and by including the same moderating variables (prior knowledge, WMC, shifting, and fluid intelligence). Schwaighofer and colleagues assumed differential effects of all investigated cognitive functions and were surprised by the null finding regarding WMC. The authors described that learners in their study were not under time pressure and thus might have offloaded WMC demands, for example by re-reading (de Jong, 2010) and proposed the lack of time pressure in their study as potential reason for not detecting the assumed influence of individual differences in WMC. We extended our replication study and added a time pressure manipulation to address to what extent individual differences in WMC matter for instructionally supported statistical problem-solving.

Problem-Solving With Worked Examples

Per definition, we encounter a problem when we want to achieve a goal but face an obstacle trying to get from our current state to the goal state (Jonassen, 2000). This holds for a novice as well as for an expert but what specifically posits a challenge varies as a function of expertise. Consequently, problem-solving is a question of distance between the current state and the desired state. This distance is, from an information processing perspective, reduced through understanding and search processes. Understanding the problem creates a

representation of the problem, referred to as problem space. Within this space, the problem solver searches for operators and by applying operators transforms a current state into a new, closer-to-the-goal state until the goal is reached (Newell & Simon, 1972). A statistical problem might exist when a researcher is curious about the relative effect of motivation and intelligence on mathematics learning but does not yet know how to investigate such a question. Statistical concepts, such as independent, dependent, or control variables, research designs, or statistical analyses act as operators, which the learner can use to solve the problem.

Searching the problem space until one finds the appropriate operator demands learners' cognitive abilities (Lee & Anderson, 2013). While searching, the learner is required to hold the current state, the desired state, and possible operators present in working memory. Because WMC is limited, problem-solving can interfere with learning. The problem solver uses mental capacity for search processes and not for schema building processes, the processes that actually lead to long-term learning (Sweller, 1988). Learning can be fostered by instructional support like worked examples, which present learners not only a problem but also the problem's solution and the steps required to reach it (van Gog & Rummel, 2010). Through this design, the learner is not tied up in search processes, experiences less cognitive load, and can focus on acquiring schemas (Sweller, 1988). It is emphasized elsewhere that worked examples foster schema acquisition by helping learners understand domain principles. Whereas learners mainly apply weak strategies (general operators such as: finding a statistical analysis) in problem-solving, they apply specific operators with worked examples (using a multiple linear regression) which contributes to acquisition of domain specific schemas (Renkl, 2014).

Moderators of the Worked Example Effect

Prior Knowledge

To be more effective than problem-solving, worked examples have to be used in early stages of cognitive skill acquisition. Learners with higher prior knowledge actually benefit more from engaging in problem-solving than in worked example study, a finding termed the expertise-reversal effect (Kalyuga, 2007; Kalyuga, Ayres, Chandler, & Sweller, 2003; Kalyuga, Rikers, & Paas, 2012). In early phases of learning, learners depend on analogy and understanding of abstract rules. In later stages, when learners have higher prior knowledge in a domain, they automatize and optimize their problem-solving ability (ACT-R framework, Anderson, Fincham, & Douglass, 1997). Worked examples aptly match the need of learners to understand underlying problem structures and domain concepts. As this need changes with increasing expertise, worked examples become less effective and eventually dysfunctional. Worked examples interfere with the maybe-not-yet-perfect but already automated problem-solving of more expert learners whose need to perfect their ability is more fittingly supported by practicing problem-solving (Atkinson, Derry, Renkl, & Wortham, 2000).

Working Memory Capacity and Shifting Ability

Working memory is a mental storage system that processes information and constantly monitors the status of this information to only maintain the most relevant information needed to achieve a goal (Miyake et al., 2000). It is often operationalized in terms of capacity because the information that can be processed and maintained simultaneously is limited (Draheim, Hicks, & Engle, 2016). Shifting refers to the ability to switch quickly between executing different tasks (Miyake et al., 2000). For example, reading a paper, switching to answer a question of a colleague in an instant, and then quickly going back to reading. Shifting requires thus a conscious switch of attention between different mental task sets which differentiates it from attentional shifts in visual processing (Miyake et

al., 2000). WMC and shifting belong to a set of basic cognitive functions referred to as executive functions. They uniquely contribute to cognition (Miyake & Friedman, 2012) but are also assumed to be critical for one another, as each ability requires aspects of the other one. For example, shifting from reading a paper to answering a colleague's question requires representation of new information in working memory. The information related to the colleague's question, currently activated in working memory, has to be replaced by information relevant to reading the paper when shifting back to the previous activity of reading (for more details on the conceptual and empirical relationship see for example Draheim et al., 2016).

In cognitive load theory, WMC serves as the central explanation as to why certain instruction is effective. Specifically, the theory states that instruction is effective when designed to reduce cognitive load, as in the case of worked examples (Sweller, 1988, 2010). Because worked examples are often associated with reduced cognitive load (e.g., Nievelstein, van Gog, Boshuizen, & Prins, 2010; Paas, 1992; van Gog, Paas, & van Merriënboer, 2006), van Gog and Rummel (2010) suggested that studying examples "may be even more effective for learners who have lower working memory capacity" (p. 160). Results of Schwaighofer and colleagues' (2016) study are interesting in this respect because worked examples were more effective than problem-solving irrespective of learner's WMC. Cognitive load theory not only emphasizes limited WMC but also that the capacity function of working memory is relative to the knowledge structures in long-term memory (LTM) (Sweller, 1988). This aligns with the proposal that knowledge in a domain allows for an extended use of WMC referred to as long-term working memory (LT-WM) (Ericsson & Kintsch, 1995). The idea of LT-WM developed from research showing that expert chess players have better recall than novices, but only for realistic and not random chess positions, and while both experts and novices recall chess positions in chunks, experts are able to store larger chunks (e.g., Chase & Simon,

1973; Ericsson & Kintsch, 1995). That is, more advanced learners likely use their low capacity more efficiently because they can rely on existing knowledge structures that novice learners have not yet developed. Individual differences in WMC might consequently not matter for the worked example effect as differences are possibly compensated by prior knowledge. Further, cognitive load theory assumes that any learner has insufficient WMC when presented with a novel, complex (high element interactivity) problems (Paas & Sweller, 2014; Sweller et al., 2019), like those used in Schwaighofer et al.'s study. Assuming individual differences in WMC matter, Schwaighofer et al. (2016) offered a different explanation for their finding that WMC did not moderate the worked example effect: The authors discussed the possibility that learners in their study found ways to "offload" cognitive load (de Jong, 2010). Because learners were not under time pressure, they may have for example gone back to the problem description to refresh relevant information in working memory whenever they needed to. It remains thus open whether the worked example effect is contingent on WMC when learners have no options to offload cognitive load.

Schwaighofer et al.'s (2016) study provided novel insights regarding other components of cognitive architecture. Learners with lower shifting ability benefited more from worked examples than from problem-solving but this benefit of worked examples over problem-solving decreased with increasing shifting ability, suggesting that it eventually reverses. In the problem-solving process, learners frequently have to switch between processing the problem description, encoding information from the materials, and applying encoded information to the problem at hand in order to reach the solution (Bassok & Novick, 2012). Learners thus shift their attention between external sources of information and further between different cognitive actions such as encoding information when making sense of the learning material and applying knowledge when generating the problem solution. A worked example might reduce the demand on learner's shifting ability as it combines the relevant

information from problem and materials which reduces the need to externally switch between problem and solution materials. Moreover, solution steps of the worked example show how problem and conceptual information relate. Hence, worked examples reduce the demand to frequently switch between mental actions of encoding and applying which makes it easier for the learner to focus on processing the relations.

The theoretical accounts for why shifting attention between different tasks or actions comes with a cost, further explain why worked examples are particularly effective for learners with lower shifting ability. Reconfiguration theory assumes that the cognitive system must reset itself to enable a person to go back to what they have been doing before attention was shifted to a different task and that costs arise due to this reconfiguration (Rogers & Monsell, 1995). In comparison to problem-solving, a worked example might support learners in returning back to the mental state they were in before switching attention as its straight forward structure of solution steps guides learners attention. Interference theory assumes that the costs of switching back to a previous activity stem from inhibiting the previous, now irrelevant activity (Alport, Styles, & Hsieh, 1994). The step-wise problem solution of a worked example directly presents the next step to the learner, making it easier to inhibit the previous activity and move on to the next operator. Without the worked example, the learner depends much more on own effort to proactively disengage from the previous activity. While it is still debated whether reconfiguration or interference theory better accounts for shifting costs, there is agreement that the theories are not mutually exclusive (Draheim et al., 2016). Both theories provide insight into how worked examples might reduce task shifting demands.

Fluid intelligence

Fluid intelligence comprises our ability to reason; to use logic and recognize patterns in new situations (König, Bühner, & Murling, 2005). Although fluid intelligence and executive functions are related to a certain degree, intelligence and executive functions are

empirically separable (Friedman et al., 2006) and each explain unique variance in learning outcomes (Yeniad et al., 2013). Empirical evidence showed that learners with lower intelligence do better in more structured instructional settings and learners with higher intelligence do better in less structured instructional settings (Snow & Lohman, 1984). Similarly, Schwaighofer et al.'s (2016) study showed that worked examples were more effective for learners with lower fluid intelligence but this difference decreased with increasing fluid intelligence suggesting that with higher fluid intelligence problem-solving becomes more effective than worked example study. Because worked examples integrate relevant information from problem and solution materials, learners are less demanded to reason which is the critical information from the problem, which are the relevant operators, and which operators are applicable to which information. Taken together, it seems likely that effects of certain support tools depend on the fluid intelligence of a learner.

Rational for Investigating Differential Effects of Instructional Support

A key ingredient to effective instruction is designing support tools that consider human cognitive architecture (Sweller, 1988, 2010). Whereas the general structure of human cognitive architecture seems fairly universal, the extent to which individuals differ on single cognitive abilities has been found to be substantial (e.g., Friedman & Miyake, 2017). Although designed considering human cognitive architecture, learning experiences or instructional support may consequently still be ineffective. According to aptitude theory, performance will even be impaired if the "inner" and "outer" environment are inapt (Snow & Lohman, 1984). Thus, aligning design of instruction with individual learner needs is desirable but necessitates knowledge of which aptitudes the design to align to.

There is considerable empirical evidence that prior knowledge should be factored in when making instructional decisions. Levels of prior knowledge determine whether a worked example or problem-solving activity is better suited (Kalyuga et al., 2003) or whether self-

explanation prompts are effective (Leppink, Broers, Imbos, van der Vleuten, & Berger, 2012). Not nearly as much is known about individual differences in executive functions or fluid intelligence as is known about prior knowledge in this respect. As executive functions and fluid intelligence matter for learning (e.g., Yeniad et al., 2013; Yuan, Steedle, Shavelson, Alonzo, & Oppezzo, 2006) and learners differ on these aptitudes (Friedman & Miyake, 2017), differential effects of instructional support seem likely and these aptitudes worth investigating.

While identifying aptitude-treatment interactions allows for more customized and probably more effective instruction, the impact of such investigations goes clearly beyond the direct practical one. The ability to describe so-called compound effects ("effects that alter the effect of other instructional effects", Sweller et al., 2019, p. 10) has most recently been emphasized as sign of matured instructional theory. Aptitude-treatment research not only pinpoints the aptitudes that matter but also limitations of effects generally assumed to be effective which results in well-described compound effects.

The Present Study

The worked example effect is backed up with convincing theoretical and empirical evidence. Although being the central component of cognitive load theory, only a few studies included measures of WMC as control (Berends & van Lieshout, 2009; van Gerven, Paas, van Merriënboer, & Schmidt, 2002) or moderating variable (Lusk et al., 2009; Schwaighofer, Vogel, et al., 2017; Seufert, Schütze, & Brünken, 2009). Schwaighofer et al.'s (2016) study utilized objective measures and tested the worked example effect's contingency on WMC that was suggested by van Gog and Rummel (2010). As the results indicate that shifting ability might be more important than WMC in instructionally supported learning, Schwaighofer et al.'s study underlined the benefit of broadening the perspective on human cognitive architecture. We replicated Schwaighofer and colleagues' study to validate this

novel perspective on how core components of our cognitive architecture influence instructional effectiveness. As in the original study, we compared the effect of worked examples and problem-solving on application-oriented knowledge which the authors defined as "knowledge necessary to identify relevant aspects of a problem as well as knowledge that are applicable to solve the problem" (Schwaighofer et al., 2016, p. 982). We further included the same moderator variables that were investigated in the original study: prior knowledge, WMC, shifting, and fluid intelligence. Schwaighofer et al. (2016) took an individual differences perspective and assumed that worked examples would be more effective than problem-solving for learners with lower WMC, but that learners with higher WMC could readily use problem-solving. In contrast, cognitive load theory assumes that (in comparison to prior knowledge) the impact of individual differences in WMC is too marginal to matter when learning complex materials (Paas & Sweller, 2014; Sweller et al., 2019). We extended the original study by including a time pressure manipulation, investigating whether WMC moderates the worked example effect when learners are not able to offload memory demands. This extension condition will contribute to answering the question whether individual differences in WMC play a role in instructionally supported learning.

Replication and pre-registration

For clear communication of the replicated and the newly investigated aspects, we distinguish the "replication" and the "extension" condition (Schmidt, 2009). The replication part of our study can be categorized as "direct" or "close" because we have repeated the experimental procedure of the original study and used the same or only slightly different measures and materials (Brandt et al., 2014; Schmidt, 2009). The study was pre-registered through the Open Science Framework (OSF) (Bichler et al., 2019, October 5).[1]

[1] Please note that Matthias Schwaighofer originally pre-registered. Matthias Schwaighofer abruptly passed away on June 10th, 2017. We created a fork of the original pre-registration, which can now be found under the first authors name.

Hypotheses

Hypotheses addressed in the replication part of our study are exactly the same as the hypotheses addressed in the original study:

Hypothesis 1: Learners acquire more application-oriented knowledge with worked examples than through problem-solving.

Hypothesis 2: Learners indicate less cognitive load when they study worked examples compared to when they problem-solve.

Hypothesis 3: Worked examples are more effective than problem-solving for learners with lower prior knowledge but problem-solving is more effective than worked examples for learners with higher prior knowledge to acquire application-oriented knowledge.

Hypothesis 4a: The benefit of worked examples over problem-solving is greater for learners with lower than for learners with higher WMC.

Hypothesis 4b: The benefit of worked examples over problem-solving is greater for learners with lower than for learners with higher shifting ability.

Hypothesis 4c: The benefit of worked examples over problem-solving is greater for learners with lower than for learners with higher fluid intelligence.

Hypotheses addressed in the extension part of our study, were not addressed in the original study:

Hypothesis 1: Learners who are under time pressure indicate higher cognitive load than learners who are not under time pressure.

Hypothesis 2: The benefit of worked examples over problem-solving is greater for learners with lower WMC than for learners with higher WMC *only* if learners are under time pressure (moderated moderation).

Method

Sample

One hundred and fifteen higher education students of social sciences programs at a German university participated in the replication condition of the study. One hundred and sixteen students participated in the extension condition (N_{Total} = 231). Participants were recruited through announcements in lectures, postings, social media, and an official university email service and compensated with 55€ or a participation certificate (needed by psychology students to complete undergraduate psychology). Participants were excluded from the sample if they had missing data (e.g., participated in only one of the two sessions). Sample characteristics are summarized in Table 1. Mean age and gender distribution of the replication condition are similar to the original sample (compare Schwaighofer et al., 2016).

Design

This study was pre-registered as laboratory study with a 2X2 between-subjects design. A repeated measure was used to assess prior and posttest knowledge, moderator variables include prior knowledge, WMC, shifting, and fluid intelligence. "Research Randomizer" (www.randomizer.org) was used to randomly allocate participants to conditions. The independent variable of the replication is worked examples: with (n_1 = 57) vs. without (n_2 = 58), where without worked examples refers to problem-solving in this paper. The second independent variable pertains to the extension and is time pressure: with (n_3 = 58) vs. without (n_4 = 58).

Material

Participants worked on six statistical problems, each of which described a hypothetical research project. Participants were asked to address the research question statistically and prompted to justify their answer (Figure 1) using textbook material on the general linear model and research methodology.

The expected solution included correctly identifying variables as independent, dependent, or control, their scale level, a research design, a statistical analysis, and statistical assumptions related to the chosen analysis. Material was presented on PowerPoint slides on a laptop. Scrolling between slides was not restricted, however, participants were asked to only go to the next problem once they solved the previous one completely.

Worked Examples and Time Pressure

The worked example showed participants how to solve the problem step by step. First step: Identifying independent and dependent variables and proposing a design for the study; Second step: Deciding which statistical analysis was appropriate; Third step: Pointing out which statistical assumptions had to be tested. Each step was shown on one slide with the step as heading and the problem as body of the slide. The solution was shown in boxes next to the problem on the right side of the slide. Solution relevant information from the problem was linked via arrows to the box the solution was presented in (Figure 2). The three worked example slides were *not* available in the problem-solving condition. Participants were prompted to explain their solutions on a sheet of paper to make sure problems were actively processed in both conditions (Schwonke et al., 2009).

Time pressure was induced by explicitly telling participants to finish all six practice problems in 45 minutes. To maintain time pressure during the intervention, a red timer counting down was shown on their desktops. Participants in the condition without time pressure also had 45 minutes to solve all problems but were not told in advance and did not see the countdown.

Measures

Knowledge tests. Application-oriented knowledge was assessed at pre- and posttest. Two parallel tests were used, both consisted of six open response items. Each item described a statistical problem similar to the practice problems but shorter (Figure 3). Participants were

asked to identify variables, suggest a design and statistical method, and name two statistical assumptions. From pre- to posttest, surface information of the problems was changed but their structure was kept the same. For example, the research question was about flavors of yogurt in pre-, and about the effectiveness of diets in posttest (surface), but both problems included a three-level independent variable, repeated measurement and an interval-scaled dependent variable (structure). Tests were coded with rubrics and points were awarded for information that was present in participants' answers. The maximum score was 26 points. The Kuder-Richardson-20 index was used to estimate reliability and showed r_{tt} = .78 for pre- and r_{tt} = .75 for posttest (N = 231, respectively).

Cognitive load. Cognitive load was measured with a 9-point rating scale developed by Paas (1992), which was translated to German by the first author of the original study. Answer options ranged from *very, very low mental effort (1) to very, very high mental effort (9).* Participants had to indicate their perceived mental effort after each statistical problem for as many problems as they completed. The average cognitive load across the total number of completed problems was used as indicator of cognitive load.

Working memory capacity. WMC was measured with the shortened version of the automated operation, reading, and symmetry span tasks (Oswald, McAbee, Redick, & Hambrick, 2015) in E-Prime Version 2.0.10.356 (Psychology Software Tools, Pittsburgh, PA).

The three automated span tasks differ only concerning the stimulus material. Participants have to indicate whether a simple mathematical equation is right or wrong (operation span), a sentence makes sense or not (reading span), or a pattern in an 8X8 matrix is symmetrical or not (symmetry span). Then, participants have to memorize a letter (operation and reading span) or the position of a red square in a 4X4 matrix (symmetry span). After a number of processing and memorization sequences, participants have to recall the

letters or the pattern of red squares in the right order. Correctly recalled elements across all six test trials are summed up and divided by the maximum possible score (30 for operation and reading, 24 for symmetry span) (Conway et al., 2005). The mean across the three tasks was used as capacity indicator.

Internal consistency for each of the three tasks was: $\alpha = .56$ (operation span, N = 228), $\alpha = .65$ (reading span, N = 230), and $\alpha = .43$ (symmetry span, N = 229) (method of Kane et al., 2004).[2]

Shifting. Shifting was measured with the computerized number-letter, color-shape, and category-switch tasks (e.g., Friedman et al., 2016; Friedman et al., 2008) using the method of Friedman et al. (2016).

In all shifting tasks there are two rules and two different stimuli indicate which rule has to be applied. In the number letter task, a number-letter pair (e.g., 4E) appears in a 2X2 matrix. When the pair appears in one of the upper two quadrants, participants have to indicate whether the number is odd or even by pressing the 'D' (odd) or 'L' (even) key. When the pair appears in one of the bottom quadrants, participants have to indicate whether the letter is a vowel or consonant by pressing the 'D' (consonant) or 'L' (vowel) key. If the rule changes between trials, it is a switch trial, if the rule stays the same it is a non-switch trial. Switch-costs are calculated by subtracting the mean reaction time of non-switch from the mean reaction time of switch trials. Higher switch-costs reflect lower shifting ability. The mean of all three tasks' switch costs was used as indicator for shifting ability.

Data was trimmed as described by Friedman et al. (2008) to handle outliers and improve normality. The split-half reliabilities (Guttman) were $r_{tt} = .88$ (number-letter), $r_{tt} = .80$ (color-shape), $r_{tt} = .81$ (category-switch), N = 231 respectively.

[2] Five participants had missing data on at least one of the three WMC tasks leading to different sample sizes in reliability analyses.

Fluid intelligence. Fluid intelligence was measured with three subtests of the computerized adaptive intelligence structure battery (INSBAT; Arendasy et al., 2012). For each subtest, the number of tasks depended on the performance of the participants.

The subtests were numerical inductive, figural inductive, and verbal deductive reasoning. In numerical inductive reasoning, the rule underlying a series of numbers has to be identified to complete the series. In figural inductive reasoning, participants see a 3X3 matrix with one empty field. They have to identify the rule to choose the correct symbol out of 8 possible symbols to complete the matrix. In verbal deductive reasoning, participants read two statements and have 45 seconds to draw a conclusion from these statements, choosing one of five possible answers. The results of all subtests were transformed into a raw score for fluid intelligence by the testing system. The reliability of each subtest was preset to $\alpha = .70$.

Control variables. Motivation was measured with the Questionnaire of current Motivation (QCM; Rheinberg, Vollmeyer, & Burns, 2001). The QCM consists of four subscales: Interest (5 items), challenge (4 items), probability of success (4 items), and anxiety (5 items) that are all answered on a 7-point Likert scale ranging from (1) disagree to (7) agree with only the extremes showing verbal anchors. See Table A1 for example items and reliabilities. Age and number of semesters were collected with other demographic data with a questionnaire.

Manipulation check. Subjective time pressure was assessed with the item "Indicate how much time pressure you felt during working on the statistical problems." Participants rated their subjective time pressure on a 10-point scale from "no time pressure" to "very, very, high time pressure" with verbal anchors for each point.

Procedure

In the first session, participants provided demographic data, gave informed consent, and completed the 35 minutes pretest. Subsequently, participants completed the executive

functions tasks and the fluid intelligence test. The first session took about 3 hours; participants were allowed to take short breaks. The second session consisted of the intervention and posttest. Participants filled out the QCM items, then worked on solving six statistical problems in either one of the four conditions for 45 minutes. Cognitive load was measured throughout the intervention after completion of each problem. After the intervention participants worked on the posttest for approx. 35 minutes.

Methodological differences between original and replication study. In the original study, the worked example was shown in plain text on three consecutive PowerPoint slides after the problem description slide. Each slide contained one solution step and its solution. In the replication study, each solution step and its solution were shown *next* to the problem description on a PowerPoint slide respectively on three consecutive slides. Relevant information was highlighted and linked with the solution via arrows. Because this change is visual in nature we believe that it has no other effect than possibly enhancing the worked example effect.

Although conceptual and application-oriented knowledge were assessed in the original study, moderation analyses were only carried out for application-oriented knowledge (Schwaighofer et al., 2016). As the moderation analyses address the main questions of the replication, we only assessed application-oriented knowledge.

To assess application-oriented knowledge more reliably, four new test items were added to the two used in the original study. Test time was adjusted from 20 minutes in the original study to 35 minutes in the replication to give participants sufficient time to solve all six problems. The intervention was 60 minutes long in the original study but participants did not need nearly that much time. Thus, the intervention phase was shortened to 45 minutes and three new practice problems were added to the three from the original study in the replication.

Differences between the pre-registration and our methodology. Additional preregistered hypotheses referring to new research questions including perceptual speed and complex problem-solving skills were not addressed in this report. Instead of the preregistered factor scores, we used mean scores across the three tasks for shifting and WMC because mean scores were used in the original study. Replication studies should not change how variables are operationalized to ensure unconfounded interpretation of replication results (LeBel, Vanpaemel, Cheung, & Campbell, 2019).

Statistical analyses. The same cut-off values for effect sizes as in the original study were applied. We set an alpha level of 5% in all analyses and we report 95% bootstrapped (5000 samples) confidence intervals for unstandardized regression coefficients in moderation analyses, which were calculated with the SPSS (IBM SPSS Statistics, Version 24) macro PROCESS Version 3 (Hayes, 2018). We controlled influence of covariates on moderator and outcome and estimated heteroscedasticity-consistent standard errors (HC3). An alpha level of 5% was used to determine statistical significance in all other analyses which are mentioned in the respective section of the addressed hypotheses. Preliminary and main analyses for the replication part of the study reported first and refer to N = 115 participants in the replication condition. Extension condition analyses refer to the full sample of N = 231 including participants from all four conditions.[3]

Results

Preliminary Analyses

Meaningful correlations and differences at pretest. WMC and fluid intelligence were highly correlated $r(114) = .48$, $p < .001$, thus WMC was controlled for in the moderation analysis for fluid intelligence and vice versa. A similar correlation was found in

[3] Five participants had missing data on at least one of the three WMC tasks leading to smaller sample sizes in analyses including WMC as moderator or control variable.

the original study.

Cognitive load was not significantly correlated with WMC $r(114) = .16$, $p = .087$, which is consistent with the original finding. Cognitive load was not correlated with fluid intelligence $r(115) = .10$, $p = .301$ or prior knowledge $r(115) = -.11$, $p = .226$, which is inconsistent with the original study. Cognitive load was correlated with the two sub-scales of QCM anxiety and challenge. Learners who experienced more cognitive load had higher anxiety $r(115) = .36$, $p < .001$ and perceived the tasks as more challenging $r(115) = .23$, $p = .015$. Thus, anxiety and challenge were used as covariates when testing for differences in cognitive load between conditions (Hypothesis 2). Similar correlations were found in the original study.

Differences in number of semester and age between groups at pretest were tested with *t*-tests. Number of semesters did not differ significantly between experimental conditions prior to the study $t(113) = .44$, $p_{two\text{-}tailed} = .661$, $d = .08$, whereas age was significantly higher in the condition with worked examples than in the condition without ($M = 24.21$, $SD = 6.23$ vs. $M = 22.49$, $SD = 3.29$), $t(84.60) = -1.84$, $p_{two\text{-}tailed} = .069$, $d = .35$ (equal variances not assumed).

Prior knowledge and knowledge at posttest were strongly correlated $r(115) = .64$, $p < .001$. Semester was correlated with prior knowledge $r(115) = .36$, $p < .001$ and posttest knowledge $r(115) = .34$, $p < .001$ but not with knowledge gains (post-pre) $r(115) = .09$, $p = .362$. Age, however, was correlated with knowledge gains $r(115) = -.21$, $p = .024$ but not with prior or posttest knowledge ($r(115) = .06$, $p = .552$ and $r(115) = -.12$, $p = .195$, respectively). Thus, age and semester were used as covariates to test Hypotheses 1 and 2 and in all moderation analyses (Hypothesis 3, 4a-c). In the original study, age and semester were controlled in the same analyses.

For descriptive statistics of variables in the replication condition see Table B1.

Worked example effect on knowledge and cognitive load (Hypotheses 1 and 2)

Descriptive statistics for prior, post-test knowledge, knowledge gains, and cognitive load *by experimental condition* are found in Table B2.

A two-factorial analysis of covariance (ANCOVA) with repeated measures including semester and age as covariates was used to test whether knowledge gains from pre- to posttest are greater in the worked examples compared to the problem-solving condition. Factor 1 was time of measurement (pre- vs. posttest) and factor 2 was worked examples (with vs. without). A statistically significant effect of time of measurement $F(1,111) = 29.86$, $p < .001$, $\eta^2 = .21$ indicated that all participants improved from pre-test to post-test irrespective of the condition they were in. A statistically significant interaction effect $F(1,111) = 4.50$, $p = .036$, $\eta^2 = .04$ (Figure 4) however showed that average knowledge gains from pre-test to posttest were greater for participants who had worked examples ($M_{pre} = 4.01$ to $M_{post} = 8.41$, $M_{gain} = 4.40$) in comparison to participants who engaged in problem-solving ($M_{pre} = 4.75$ to $M_{post} = 8.16$, $M_{gain} = 3.59$). This interaction effect was also statistically significant in the original study. Thus, worked examples were more effective than problem-solving in both studies.

Differences in cognitive load were tested with a one-factorial univariate ANCOVA with worked examples (with vs. without) as independent and cognitive load as dependent variable. There was no statistically significant difference between worked examples and problem-solving in cognitive load $F(1,111) = 0.30$, $p = .585$, $\eta^2 = .003$. This finding aligns with the original study.

Prior Knowledge Moderation (Hypothesis 3)

A moderation analysis (Hayes, 2018) with worked examples (with vs. without) as independent, knowledge gains (post-test – pre-test score) as dependent, and prior knowledge as moderating variable was used to test whether the worked example effect was contingent on

levels of prior knowledge. The unstandardized regression coefficient for the conditional effect of worked examples on knowledge gains was significantly different from 0: b = -0.38, p = .015, 95% $CI_{5000boot}$ [-0.672, -0.095]. Simple slope analysis showed the following effects on different levels of prior knowledge: $M_{(-\ 1SD)}$ = 1.58, b = 1.99, SE = 0.67; M = 4.29, b = 0.95, SE = 0.48; $M_{(+\ 1SD)}$ = 7.01, b = -0.09, SE = 0.68. Learners with lower prior knowledge acquired more knowledge in the condition with worked examples, learners with higher prior knowledge, in contrast, acquired more knowledge in the condition with problem-solving (see Figure 5 for a visualization of the interaction effect). Prior knowledge did not significantly moderate the worked example effect in the original study.

Executive Functions and Fluid Intelligence Moderations (Hypotheses 4a, 4b, & 4c)

A moderation analysis with worked examples (with vs. without) as independent, knowledge gains (posttest – pretest score) as dependent, and WMC as moderating variable was used to test whether the worked example effect was contingent on levels of WMC. WMC did not moderate the effect of worked examples on knowledge gains b = 1.47, p = .346, 95% $CI_{5000boot}$ [-4.814, 3.525], (N = 114). Thus, the worked example effect is not contingent on levels of WMC, learners with lower and higher WMC benefited from worked examples over problem-solving. WMC did not moderate the worked example effect in the original study.

The moderation analysis including shifting ability as moderator showed a significant interaction effect of worked examples (with vs. without) and shifting ability on knowledge gains b = 0.007, p = .033, 95% $CI_{5000boot}$ [0.002, 0.014]. Note that switch costs were not recoded and that *higher* values indicate a *lower* shifting ability, that is, an increase in scores reflects a decrease in ability. Thus, with every unit (millisecond) shifting ability decreases, the difference between worked examples and problem-solving increases by .007 points in the test. That means, the benefit of worked examples over problem-solving increases with

decreasing shifting ability. Simple slope analysis showed the following effects on different levels of shifting: $M_{(-1SD)}$ = 37.88, b = -0.04, SE = 0.74; M = 183.77, b = 1.05, SE = 0.51; $M_{(+1SD)}$ = 329.66, b = 2.13, SE = 0.81. Please see Figure 6 for a visualization. This moderation effect was found in the original study as well.

The moderation analysis including fluid intelligence as moderator showed that the worked example effect on knowledge gains was not contingent on levels of fluid intelligence b = -0.01, p = .493, 95% $CI_{5000boot}$ [-0.364, 1.115] (N = 114). Learners with lower and higher fluid intelligence benefited from worked examples over problem-solving. In the original study, worked examples were more beneficial for learners with lower than for learners with higher fluid intelligence, indicating that with increasing fluid intelligence problem-solving becomes more effective.

Extension Hypotheses

Manipulation check. Whether learners in the condition with time pressure in fact felt more time pressure was tested with an independent *t*-Test with time pressure as independent and subjective time pressure as dependent variable. Learners in the time pressure condition subjectively felt under higher time pressure (M = 5.00, SD = 2.25) than learners in the no time pressure condition (M = 3.00, SD = 2.19). This difference was statistically significant $t(229)$= 6.85, p < .001, d = .9. Thus, the manipulation was successful.

Effect of time pressure on cognitive load (Hypothesis 1). Time pressure and worked examples were included in a 2-factorial ANCOVA as independent variables with cognitive load as dependent variable controlling for age, semester, and the QCM scales anxiety and challenge. Only anxiety and challenge were significant covariates. The main effect of time pressure on cognitive load was significant $F(1,222) = 5.43$, p = .021, η^2 = .02, learners in the time pressure condition indicated higher cognitive load (M = 6.41, SD = 1.20) compared to learners in the no time pressure condition (M = 6.05, SD = 1.22). The main

effect of worked examples on cognitive load was not statistically significant $F(1,222) = 1.61$, $p = .206$, $\eta^2 = .01$, indicating that there was no difference between worked examples and problem-solving in cognitive load. The interaction effect of time pressure and worked example was also not statistically significant $F(1,222) = 0.03$, $p = .864$, $\eta^2 < .001$. This means, only time pressure but not the study method (worked examples vs. problem-solving) nor any of the study methods in combination with time pressure affected how much cognitive load learners experienced.

Three-way interaction of time pressure, working memory capacity, and worked examples (Hypothesis 2). A moderated moderation analysis with worked examples as independent variable, WMC as primary and time pressure as secondary moderator with knowledge acquisition as dependent variable was conducted (Figure 7). The three-way interaction effect was not statistically significant $b = -0.85$, $p = .434$, 95% $CI_{5000boot}$ [-9.113, 7.362] (N = 226). Even when under time pressure, WMC did not moderate the worked example effect on knowledge acquisition. Hence, extension Hypothesis 2 did not receive empirical support.

Discussion

We replicated Schwaighofer et al.'s (2016) study and investigated whether there are differences in application-oriented knowledge and in cognitive load between worked example study and problem-solving. All participants gained application-oriented knowledge from pre-test to post-test, however the gains were greater for participants in the worked example than in the problem-solving condition. There was no difference between worked examples and problem-solving with respect to cognitive load. Further, we investigated moderators of the worked example effect. Learners with lower prior knowledge benefited from worked examples, whereas the benefit of worked examples over problem-solving decreased with increasing prior knowledge of learners. Also, worked examples were more beneficial the

lower a learner's shifting ability was in comparison to problem-solving. In contrast, the worked example effect was not found to be contingent on levels of WMC and fluid intelligence; learners with lower and higher WMC or fluid intelligence similarly benefitted from worked examples over problem-solving. We extended Schwaighofer et al.'s study and included a time pressure manipulation. Participants in the time pressure condition indicated higher cognitive load than participants who were not under time pressure. With respect to our moderated moderation hypothesis, we found that even when under time pressure, worked examples were equally beneficial for learners with lower and higher WMC.

We follow the suggestion of LeBel et al. (2019) on nuanced statistical language for interpretation of replication results in our discussion. The authors suggest to use *signal-consistent* and *signal-inconsistent* for statistically significant replication findings. *Consistent* is used when the replication confidence interval (CI) includes the original effect size (ES) and *inconsistent* is used when the replication CI excludes the original ES. *Signal-inconsistent* is further differentiated to express whether the replication ES was larger than, smaller than, or in the opposite direction of the original ES. *No signal-consistent* and *no signal-inconsistent* is used for statistically non-significant replication findings and to indicate whether the replication CI includes or excludes the original ES. Unfortunately, above mentioned terms were also suggested to express replication results with respect to statistically non-significant effects in the original study (LeBel et al., 2019). To avoid confusion, we recommend *no signal repeated – consistent* and *inconsistent* for cases where neither the original nor the replication study detected an effect, with *consistent* indicating that the ES of the original study was included in the replication CI, and *inconsistent* indicating that the ES of the original study was excluded in the replication CI. Further we recommend to use *no signal – signal (consistent* or *inconsistent)* for cases in which effects are statistically non-significant in the original study, but are statistically significant in the replication study. We note that

consistent and *inconsistent* cannot be interpreted when studies report unstandardized effect sizes, as is the case in our study. An overview of the replication status of effects investigated in our study is found in Table 2 which shows an *adaption* of LeBel et al.'s (2019) categorization to our case.

Worked Examples, Time Pressure, Knowledge Acquisition, and Cognitive Load

Learners who solved statistical problems with worked examples gained more knowledge compared to learners who solved problems without worked examples. Because the signal was detected consistently, we conclude that the worked example effect on knowledge acquisition was replicated. The worked example effect received empirical support in numerous studies, but mostly for well-structured and rule-based problems in domains like mathematics (van Gog & Rummel, 2010). Although investigated as a support tool for more complex skill acquisition such as argumentation (Hefter et al., 2014), worked examples had hardly been systematically investigated as support with respect to solving less rule-based problems in the domain of statistics. Worked examples effectively supported learning probability calculation or calculation of mean, mode, and median (Paas, 1992; Renkl, 1997), learning goals that require understanding rule-based operations. However, there are less defined problems in the domain of statistics that require knowledge of research methods and statistical analyses and the ability to apply this knowledge. Domain principles might be less straightforward, operators might be applied in relatively random order, and multiple solutions might be equally appropriate. Our study shows that worked examples foster application-oriented knowledge needed to solve such problems as well.

As to why worked examples are more effective than problem-solving, our results do not support the often assumed reduced cognitive load mechanism (Sweller, 1988; van Gog & Rummel, 2010). There was no statistically meaningful difference in cognitive load between worked examples and problem-solving; neither in the original nor the replication study. Our

extension analysis including time pressure showed that cognitive load was higher when learners were under time pressure but there was no interaction effect of worked examples and time pressure on cognitive load. Thus, our results are less in keeping with the idea that reduced cognitive load explains the advantage of worked examples over problem-solving, but more consistent with the explanation that worked examples foster understanding of domain concepts (Koedinger, Corbett, & Perfetti, 2010; Renkl, 2014).

Although worked examples reduced cognitive load in comparison to problem-solving during learning (e.g., Nievelstein et al., 2010; Paas & van Merriënboer, 1994; van Gog et al., 2006) or posttest performance (e.g., Paas, 1992) in a number of studies, we argue that it is more informative to know which cognitive processes the cognitive load rating of a learner reflects than to know how much cognitive load a learner reports. In our study for example, all learners invested similar amounts of mental effort, but as to which specific processes mental effort was invested remains unknown. We conclude that cognitive load ratings are a valid indicator of learner experience but these ratings do not reveal the mechanism by which worked examples are effective. The conclusion that it is more about the "kind" and less about the "quantity" of load is in line with results of other studies that found instructional effects in absence of differences in cognitive load (e.g., de Koning, Tabbers, Rikers, & Paas, 2010; Lusk & Atkinson, 2007).

As reduced cognitive load in example study conditions was found under restricted and very short study times (for example 3 mins. in van Gog et al., 2006), our findings indicate that this might not generalize to more ill-structured and complex learning settings (de Jong, 2010) in which realistic study materials and realistic study times are utilized. The subjectively rated invested mental effort might rather be indicative of how much time pressure, anxiety or how challenged learners felt by the task. Thus, based on our data, cognitive load possibly reflects metacognitive, motivational, or affective characteristics of a

learner (see for example Feldon, Callan, Juth, & Jeong, 2019 for a discussion of motivational and affective factors within cognitive load theory).

Whether cognitive load in fact reflects load in working memory has to be assessed relative to an individual's WMC (de Jong, 2010) but WMC is seldom clearly conceptualized and in even less studies measured (Anmarkrud, Andresen, & Bråten, 2019). Considering that use of WMC is assumed to be a function of prior knowledge (Sweller, 1988), cognitive load should be assessed relative to WMC and prior knowledge. As put in 1988 "any potential measure [of cognitive load] must be capable of simultaneously accounting for problem difficulty, subject knowledge, and strategy used" (Sweller, 1988, p. 263). It seems such a measure (still) does not yet exist (Anmarkrud et al., 2019). However, a question to be addressed in the future is whether such a measure, if it can be developed, provides the answers that we are seeking. Namely, whether it reveals the cognitive processes meaningful for learning and whether the amount of cognitive load or mental effort matters as long as learners engage in meaningful learning processes.

Prior Knowledge Moderation

Prior research indicates an aptitude-treatment interaction of worked examples and prior knowledge. Worked examples are helpful when prior knowledge is low but detrimental when prior knowledge is high (Kalyuga, 2007). Results of the replication align with prior research but stand in contrast to results of the original study. This can be categorized as a *no signal-signal (consistent)* scenario as the original study's effect was in the same direction as the replication study's effect. The assumed reason for the unexpected null finding in the original study was the very low variability in prior knowledge of the sample (Schwaighofer et al., 2016). We conclude that worked examples support learners with lower prior knowledge to construct internal representations but interfere with the use of already constructed internal representations in learners with higher prior knowledge (Atkinson et al., 2000; Kalyuga,

2007) as our findings align with previous research and because the authors of the original study offered a reasonable explanation for why their results differed from the expectation.

Executive Functions and Fluid Intelligence Moderations

Worked examples were hypothesized to be more effective for learners with lower WMC than for learners with higher WMC but neither the original nor the replication study provided empirical support (*no-signal repeated* replication scenario). Even when learners were under time pressure, a scenario in which demands on WMC should become unavoidable and individual differences should become visible, learners with lower and higher memory capacity benefited equally from worked examples over problem-solving.

While the results of only the original study were inconclusive with respect to whether WMC moderates the worked example effect, the repeated null finding and lack of support for the alternative time pressure explanation together increase confidence in the conclusion that individual differences in WMC do not noticeably influence the worked example effect. This pattern of findings does not support van Gog and Rummel's (2010) proposition of differential WMC effects. It rather aligns with the assumption of cognitive load theory that individual differences in WMC do not matter as all learners, including those with relatively high WMC, have insufficient processing capacity when dealing with complex materials (Paas & Sweller, 2014). It follows that in these cases, learners with higher WMC should also benefit from instructional support such as worked examples. Further, cognitive load theory has always stressed the important link of WMC and prior knowledge. More expert learners can retrieve multiple elements as one piece of information and thus utilize limited capacity much more efficiently (Sweller, 1988; Sweller et al., 2019). This is also supported by research showing that expertise in a domain allows for more efficient processing through use of the so called long-term working memory (Ericsson & Kintsch, 1995). Thus, individual differences in WMC might also be rendered meaningless for learning in contexts in which knowledge is

required or helps to solve tasks.

As both the original and our replication study indicate (*signal-consistent* replication scenario), the benefit of worked examples over problem-solving decreases with increasing shifting ability. Learners had to switch between different information sources but also between encoding and applying information to solve the problems in both studies. Because the worked example integrated relevant information and provided a step-wise solution procedure, we assume that demands on learners' shifting ability were reduced. Due to the need of reconfiguration to a previous mental state (Rogers & Monsell, 1995) and of actively disengaging from the currently active but no longer relevant mental state (Alport et al., 1994), task shifting comes at a cost. Although van Gog and Rummel's (2010) assumption of influential individual differences did not apply, in our study, to WMC, it instead seems to apply to the cognitive function of shifting. Worked examples apparently aptly fit to learners with lower shifting ability because their design provides support for reconfiguration of the cognitive system and disengagement from previous mental sets. Learners with higher shifting ability can effectively engage in these cognitive processes; they thus do not experience high shifting costs and consequently their learning during problem-solving is not impaired.

These findings expand our knowledge about how shifting influences the effectiveness of instruction and underpins the importance of this ability for learning, for which past research has already provided correlational evidence (Yeniad et al., 2013). Aligned with the suggestion of "a more dynamic and multidimensional approach to understanding working memory" (Sepp et al., 2019, p. 2), we propose that instructional theory will benefit from a broader view on cognitive architecture. Beyond greater consideration of task shifting, specifically looking deeper into how different cognitive components interact seems, based on our research, promising for advancing instructional theory. Inevitably in complex learning, learners face their WMC constraints and compensate for these by utilizing other cognitive

functions. Namely, our limited WMC makes shifting between different elements unavoidable. Those learners who are better at shifting between different elements in the material or between different mental actions would be expected to rely less on instructional support than those who are not as good at shifting (Schwaighofer et al., 2016). Interesting in this respect is also whether task shifting might be less tied to prior knowledge than WMC seems to be. While there are evidence and long-standing theoretical considerations that the use of one's WMC is inherently tied to prior knowledge (Ericsson & Kintsch, 1995; Paas & Sweller, 2014; Sweller, 1988; Sweller et al., 2019), we do not know of investigations or theoretical considerations that support a similar relationship of shifting ability and prior knowledge. In fact, Baddeley's working memory model, on which cognitive load theory is based, assumes the central executive to be a domain-general attentional control mechanism whose sub-processes are capacity to focus attention, dividing attention, and: switching attention (Baddeley, 1992, 2000; Baddeley, 2002). It would be informing to investigate if this attentional system operates independently of prior knowledge or whether schemas in long-term memory are actually involved in steering attentional resources.

With respect to the moderating role of fluid intelligence, we did not detect a signal where the original study detected a signal, a *signal-inconsistent* scenario. The original study found that the benefit of worked examples over problem-solving decreased with increasing fluid intelligence to eventually reverse (Schwaighofer et al., 2016), which is in line with earlier research on aptitude-treatment interactions (Snow & Lohman, 1984). We found that worked examples were more effective than problem-solving irrespective of fluid intelligence. Worked examples integrate relevant information, link concepts and problem information, and present all of this in a condensed form to the learner, thus reduce the reasoning demands. Schwaighofer et al. (2016) argued that with the worked example present, learners with lower fluid intelligence were better equipped to understand what information from the text book

materials were relevant. Learners with higher fluid intelligence were able to reason with the text book material even without the help of the worked example. However, not only was reasoning required to solve the statistical problems but also knowledge. The worked examples in our study might have helped to construct the needed knowledge and thus were effective for learners with lower and higher fluid intelligence equally. The inconsistent findings leave us to conclude that either explanation might hold and that future research is needed to determine which will persist. Future studies may further systematically investigate if intelligence differentially affects learning on different levels of prior knowledge (Leutner, 2002).

Limitations

The original and replication study differ in certain aspects from classical worked example studies. Typically, worked example studies use example-problem pairs (Sweller, 1988) and instructional input precedes worked example study or problem-solving (Renkl, 2014). In contrast, we used six problems subsequently in the intervention phase and instructional materials were not presented before worked example study or problem-solving but were accessible throughout the learning phase. Although it is unlikely that these differences to other studies selectively contributed to the absence of the hypothesized effects regarding WMC or cognitive load while not leading to the absence of the hypothesized worked example or shifting moderation effect, we suggest that future research utilizes a typical worked example study format.

Regarding our sample characteristics, the replication study's sample included a higher variety of study programs than the original study's sample. However, low variability in prior knowledge was the assumed reason why the prior knowledge moderation was not found in the original study, therefore we intentionally employed a sample with potentially higher variability in prior knowledge.

The reliability of the automated operation, symmetry, and reading span was relatively low which might be a reason why a WMC moderation effect was not detected. However, the reliabilities of all three tasks were high in the original study where this effect was also not found, which is why we do not consider the low reliability in our study as a likely cause for the absence of the WMC moderation.

Part of the theoretical explanation of the shifting moderation was our interpretation that learners shift less between problem and learning materials in worked example study compared to problem-solving. This could be further substantiated by analyzing log-data that provide insight into how learners specifically navigate between learning materials and whether moves between different sources are more frequent in problem-solving than worked example study. Unfortunately, our study does not provide these kind of data as it was conducted in paper-pencil format and future research is needed to follow-up on our interpretation.

Future Research

We suggest to investigate all four moderators of our study in a) different domains, b) with different kinds of instructional support, and c) with different samples to increase the credibility of the effects we have found. Studies placed in real world learning environments such as classrooms would inform us whether the observed effects hold beyond the authentic but still rather lab-based setting we used in our study. It would be of merit to include inhibition in such investigations as the ability to stay focused and deliberately suppress distractors (Miyake et al., 2000) seems an especially important feature of cognitive architecture within classroom settings. As it is fairly well established that executive functions are rather stable (Schwaighofer, Fischer, & Bühner, 2015), it would be worth investigating whether training the “entire set” of executive functions through operating at the limits of a learners cognitive capacities is possible (Schwaighofer, Bühner, & Fischer, 2017;

Schwaighofer et al., 2015). If executive functions are malleable through taxing their interplay, such training could be integrated within authentic learning experiences by ensuring that the complexity of the learning task challenges a learner's cognitive functioning adaptively, always operating at the learner's limits. Even if this would not yield improved single cognitive functions, it might improve a learner's ability to flexibly use all available cognitive functions.

Our results emphasize that aligning design of instruction with individual learner needs is desirable but it calls for more customized instruction which is challenging on many levels. As intelligent tutoring systems for example have already shown success in tailoring instruction to individual learner needs (Graesser, Hu, & Sottilare, 2018), future research could utilize technology more systematically to design instruction that is aligned with human cognitive architecture as well as flexible in adapting to different "inner" environments of learners.

Investigating the interplay of cognitive functions will certainly result in an even more adequate description of human cognitive architecture. This includes investigating how for example WMC and shifting ability interact, but also how these functions can be utilized by learners with different levels of prior knowledge. Domain-specific measures that capture this relation would provide deeper insight into this assumed interplay. As it is already difficult to assess general cognitive functions in their purest form (task-impurity problem, Miyake & Friedman, 2012), development of such measures should not be underestimated.

While instructional theory has built on cognitive psychology and used conceptualizations of cognitive architecture to understand and explain instruction and learning, not all aspects of cognitive architecture were integrated in instructional theory nor was instructional theory always adapted to updated cognitive models. For example, the central executive of Baddeley's model is barely mentioned in cognitive load theory (Schüler,

Scheiter, & van Genuchten, 2011). Current advances in instructional theory suggest a domain-general attentional resource (Sepp et al., 2019) which taken together with our findings pinpoint shifting ability as a crucial factor in instructionally supported learning, especially if this ability would turn out to be rather independent of prior knowledge. Thus, future research that incorporates so far unattended parts of recent cognitive models enables more precise instructional design implications. Moreover, cross-checking instructional theory against different cognitive, specifically working memory models, (Schüler et al., 2011) may contribute to further understand how current models of cognitive architecture look like and what they imply for learning and instruction. As multiple, partially conflicting, models of cognitive architecture exist and continue to develop (for example, Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Cowan, 2001; Engle, Tuholski, Laughlin, & Conway, 1999; Himi, Bühner, Schwaighofer, Klapetek, & Hilbert, 2019; Miyake et al., 2000; Oberauer & Kliegl, 2006; Oberauer, Süß, Wilhelm, & Wittman, 2003), instructional theory has to continuously explore and possibly integrate the of advances in cognitive psychology (Schüler et al., 2011)

Conclusion

The original and our replication study used objective measures to assess WMC, shifting ability, and fluid intelligence, and expanded knowledge about factors that influence instructional effectiveness beyond prior knowledge. Our results showed that inter-individual differences in WMC are, if at all, marginally relevant for instructionally supported learning and that in complex learning learners with lower and higher WMC benefit from worked examples over problem-solving alike. Further, we showed that task shifting is another important component of cognitive architecture that matters for effective instruction. Compared to WMC, task shifting might be less dependent on knowledge in long-term memory. We conclude that adopting a more comprehensive view on cognitive architecture

will advance our knowledge on how, why, and for whom which instructional support works. More detailed knowledge about compound effects will enable us to design more customized instruction and potentially more effective learning experiences.

References

Alport, A., Styles, E. A., & Hsieh, S. (1994). Shifting Intentional Set: Exploring the Dynamic Control of Tasks. In C. Umita & M. Moscovitch (Eds.), *Attention and performance XV* (pp. 421–452). Hillsdale, NJ: Erlbaum.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(4), 932–945.

Anmarkrud, Ø., Andresen, A., & Bråten, I. (2019). Cognitive Load and Working Memory in Multimedia Learning: Conceptual and Measurement Issues. *Educational Psychologist*, 1–23.

Arendasy, M., Hornke, L. F., Sommer, M., Häusler, J., Wagner-Menghin, M., Gittler, G., . . . Körtner, T. (2012). *Intelligenz-Struktur Batterie (INSBAT) [INtelligence Structure Battery]. Manual.* Mödling, Austria: Schuhfried GmbH.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of educational research, 70*(2), 181–214.

Baddeley, A. (1992). Working memory. *Science, 255*(5044), 556–559.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in cognitive sciences, 4*(11), 417–423.

Baddeley, A. D. (2002). Is working memory still working? *European psychologist, 7*(2), 85–97.

Bassok, M., & Novick, L. R. (2012). Problem solving. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 413–432). NY, NY: Oxford University Press.

Berends, I. E., & van Lieshout, E. C. D. M. (2009). The effect of illustrations in arithmetic problem-solving: Effects of increased cognitive load. *Learning and Instruction, 19*(4),

345–353.

Bichler, S., Bühner, M., Fischer, F., Stadler, M., & Greiff, S. (2019, October 5). Fork of Moderators of the worked example effect. Retrieved from https://osf.io/dx6qv/

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive psychology, 4*(1), 55–81.

Conway, A. R., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*(2), 163–183.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*(5), 769–786.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences, 24*(1), 87–114.

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science, 38*(2), 105–134.

de Koning, B. B., Tabbers, H. K., Rikers, R. M., & Paas, F. (2010). Attention guidance in learning from a complex animation: Seeing is understanding? *Learning and Instruction, 20*(2), 111–122.

Draheim, C., Hicks, K. L., & Engle, R. W. (2016). Combining reaction time and accuracy: The relationship between working memory capacity and task switching as a case example. *Perspectives on Psychological Science, 11*(1), 133–155.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working

memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of Experimental Psychology: General, 128*(3), 309–331.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*(2), 211–245.

Feldon, D. F., Callan, G., Juth, S., & Jeong, S. (2019). Cognitive load as motivational cost. *Educational Psychology Review*, 1–19.

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex, 86*, 186–204.

Friedman, N. P., Miyake, A., Altamirano, L. J., Corley, R. P., Young, S. E., Rhea, S. A., & Hewitt, J. K. (2016). Stability and change in executive function abilities from late adolescence to early adulthood: A longitudinal twin study. *Developmental psychology, 52*(2), 326–340.

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological science, 17*(2), 172–179.

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General, 137*(2), 201–225.

Ginns, P., & Leppink, J. (2019). Special Issue on Cognitive Load Theory: Editorial. *Educational Psychology Review*, 1–5. doi:10.1007/s10648-019-09474-4

Graesser, A. C., Hu, X., & Sottilare, R. (2018). Intelligent tutoring systems. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International handbook of the learning sciences* (pp. 246–255). New York and London: Routledge.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2nd ed.). New York, NY: Guilford

Publications.

Hefter, M. H., Berthold, K., Renkl, A., Riess, W., Schmid, S., & Fries, S. (2014). Effects of a training intervention to foster argumentation skills while processing conflicting scientific positions. *Instructional Science, 42*(6), 929–947.

Himi, S. A., Bühner, M., Schwaighofer, M., Klapetek, A., & Hilbert, S. (2019). Multitasking behavior and its related constructs: Executive functions, working memory capacity, relational integration, and divided attention. *Cognition, 189*, 275–298.

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development, 48*(4), 63–85.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*(4), 509–539.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*(1), 23–31.

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.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133*(2), 189–217.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2010). *The Knowledge-Learning-Instruction (KLI) framework: Toward bridging the science-practice chasm to enhance robust student learning*. (CMU-HCII Tech Rep. No. 10–102).

König, C. J., Bühner, M., & Murling, G. (2005). Working memory, fluid intelligence, and attention are predictors of multitasking performance, but polychronicity and

extraversion are not. *Human performance, 18*(3), 243–266.

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A Brief Guide to Evaluate Replications. *Meta-Psychology, 3*, 1–9. Retrieved from https://open.lnu.se/index.php/metapsychology/article/view/843/1835

Lee, H. S., & Anderson, J. R. (2013). Student learning: What has instruction got to do with it? *Annual review of psychology, 64*, 445–469.

Leppink, J., Broers, N. J., Imbos, T., van der Vleuten, C. P., & Berger, M. P. (2012). Self-explanation in the domain of statistics: an expertise reversal effect. *Higher Education, 63*(6), 771–785.

Leutner, D. (2002). The fuzzy relationship of intelligence and problem solving in computer simulations. *Computers in Human Behavior, 18*(6), 685–697.

Lusk, D. L., Evans, A. D., Jeffrey, T. R., Palmer, K. R., Wikstrom, C. S., & Doolittle, P. E. (2009). Multimedia learning and individual differences: Mediating the effects of working memory capacity with segmentation. *British Journal of Educational Technology, 40*(4), 636–651.

Lusk, M. M., & Atkinson, R. K. (2007). Animated pedagogical agents: Does their degree of embodiment impact learning from static or animated worked examples? *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 21*(6), 747–764.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*(6), 537–542.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions four general conclusions. *Current Directions in Psychological Science, 21*(1), 8–14.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive psychology, 41*(1), 49–100.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nievelstein, F., van Gog, T., Boshuizen, H. P. A., & Prins, F. J. (2010). Effects of conceptual knowledge and availability of information sources on law students' legal reasoning. *Instructional Science, 38*(1), 23–35.

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language, 55*(4), 601–626.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence, 31*(2), 167–193.

Oswald, F. L., McAbee, S. T., Redick, T. S., & Hambrick, D. Z. (2015). The development of a short domain-general measure of working memory capacity. *Behavior Research Methods, 47*(4), 1343–1355.

Paas, F., & Sweller, J. (2014). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (Vol. 2, pp. 27–42). New York, NY: Cambridge University Press.

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*(4), 422–434.

Paas, F. G., & van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of*

Educational Psychology, 86(1), 122–133.

Primi, R., Ferrão, M. E., & Almeida, L. S. (2010). Fluid intelligence as a predictor of learning: A longitudinal multilevel approach applied to math. *Learning and Individual Differences, 20*(5), 446–451.

Renkl, A. (1997). Learning from worked‐out examples: A study on individual differences. *Cognitive Science, 21*(1), 1−29.

Renkl, A. (2014). Toward an instructionally oriented theory of example‐based learning. *Cognitive Science, 38*(1), 1−37.

Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern-und Leistungssituationen (Langversion, 2001). *Diagnostica, 2*, 57–66.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124*(2), 207–231.

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence, 53*, 118–137.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90–100.

Schüler, A., Scheiter, K., & van Genuchten, E. (2011). The role of working memory in multimedia instruction: Is working memory working during learning from text and pictures? *Educational Psychology Review, 23*(3), 389–411.

Schwaighofer, M., Bühner, M., & Fischer, F. (2016). Executive functions as moderators of the worked example effect: When shifting is more important than working memory capacity. *Journal of Educational Psychology, 108*(7), 982–1000.

Schwaighofer, M., Bühner, M., & Fischer, F. (2017). Executive functions in the context of complex learning: Malleable moderators? *Frontline Learning Research, 5*(1), 58–75.

Schwaighofer, M., Fischer, F., & Bühner, M. (2015). Does working memory training transfer? A meta-analysis including training conditions as moderators. *Educational Psychologist, 50*(2), 138–166.

Schwaighofer, M., Vogel, F., Kollar, I., Ufer, S., Strohmaier, A., Terwedow, I., . . . Fischer, F. (2017). How to combine collaboration scripts and heuristic worked examples to foster mathematical argumentation–when working memory matters. *International Journal of Computer-Supported Collaborative Learning, 12*(3), 281–305.

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*, 258–266.

Sepp, S., Howard, S. J., Tindall-Ford, S., Agostinho, S., & Paas, F. (2019). Cognitive load theory and human movement: towards an integrated model of working memory. *Educational Psychology Review*, 1–25.

Seufert, T., Schütze, M., & Brünken, R. (2009). Memory characteristics and modality in multimedia learning: An aptitude–treatment–interaction study. *Learning and Instruction, 19*(1), 28–42.

Snow, R. E., & Lohman, D. F. (1984). Toward a theory of cognitive aptitude for learning from instruction. *Journal of Educational Psychology, 76*(3), 347–376.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*(2), 123–138.

Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional

design: 20 years later. *Educational Psychology Review*, 1–32.

van Gerven, P. W. M., Paas, F. G. W. C., van Merriënboer, J. J. G., & Schmidt, H. G. (2002). Cognitive load theory and aging: Effects of worked examples on training efficiency. *Learning and Instruction, 12*(1), 87–105.

van Gog, T., Paas, F., & van Merriënboer, J. J. G. (2006). Effects of process-oriented worked examples on troubleshooting transfer performance. *Learning and Instruction, 16*(2), 154–164.

van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review, 22*(2), 155–174.

Yeniad, N., Malda, M., Mesman, J., van IJzendoorn, M. H., & Pieper, S. (2013). Shifting ability predicts math and reading performance in children: A meta-analytical study. *Learning and Individual Differences, 23*, 1–9.

Yuan, K., Steedle, J., Shavelson, R., Alonzo, A., & Oppezzo, M. (2006). Working memory, fluid intelligence, and science learning. *Educational Research Review, 1*(2), 83–98.

Tables

Table 1
Sample characteristics

	N	M_{Age}	*(SD)*	*Women*	*Men*	*Semester Range*
Replication	115	23.34	5.03	86%	14%	11
Extension	116	22.38	3.43	88%	12%	11
Full Sample	231	22.40	4.33	87%	13%	11

Table 2
Overview of original and replication effects using nuanced language

Hypothesis	Support of hypothesis?		Replication Status of Effect
	Original Study	Replication Study	
1. Effect of worked examples on application-oriented knowledge acquisition	+	+	*Signal – consistent*
2. Effect of worked examples on cognitive load	-	-	*No signal – repeated*
3. Moderating role of prior knowledge	-	+	*No signal – signal*
4. a. Moderating role of WMC*	-	-	*No signal – repeated*
4. b. Moderating role of shifting	+	+	*Signal – (consistent)*
4. c. Moderating role of fluid intelligence	+	-	*No signal – (consistent)*

Note. * Working memory capacity. We reported *unstandardized effect sizes* for moderation analyses, which has to be considered when interpreting "consistent" and "inconsistent" as proposed by LeBel and colleagues (2019). We thus refrained from categorizing consistent and inconsistent. We used (consistent) for moderation analyses with shifting and fluid intelligence to indicate that the effects were in the same direction.

Figures

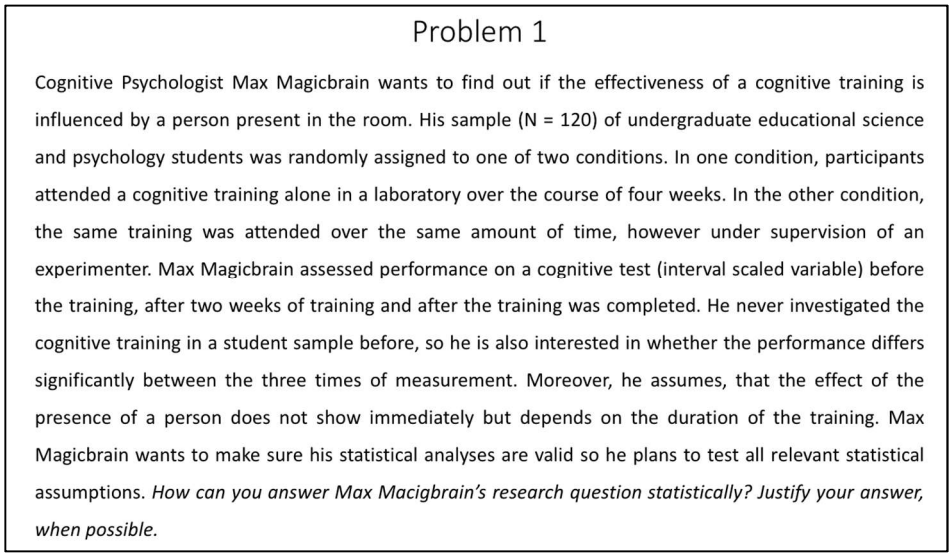

Figure 1. Example of material: first of six statistical problems that had to be solved during the intervention (translated from German).

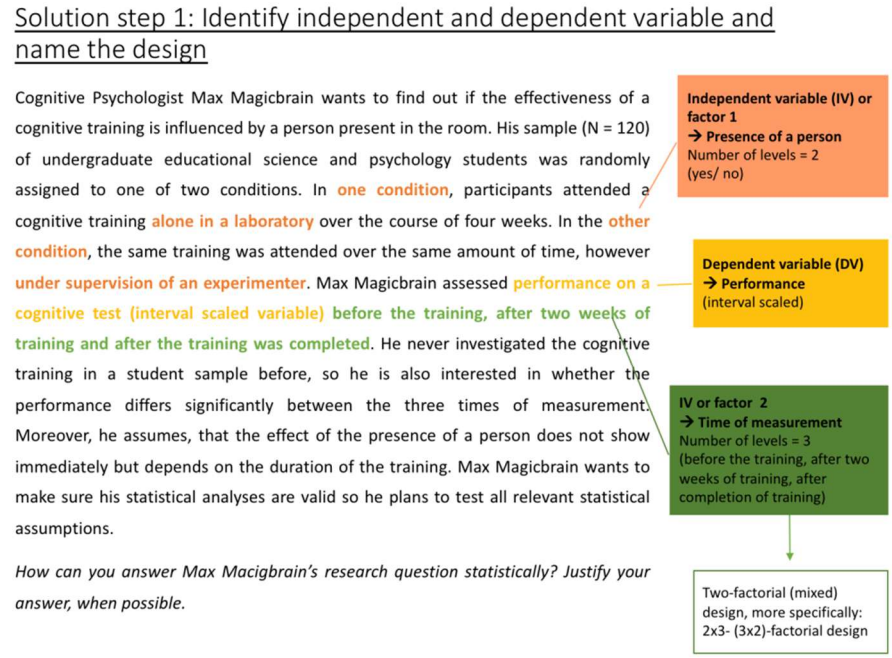

Figure 2. First of three worked example slides showing the problem, the solution step and the solution. Shown is step 1: linking the concepts of independent and dependent variable to problem information including scale level of variables. Application of textbook material on design to the specific case of the problem (translated from German).

A researcher wants to know if interest in math (interval scaled variable), prior knowledge in math (interval scaled variable), and gender (dichotomous variable) predicts how much effort students are willing to invest in studying for their math exam (interval scaled variable). The researcher wants to know in addition, which of the predictors is the strongest. *What are the independent and dependent variables? How can you answer this research question statistically? What are two assumptions that have to be met for the statistical method you choose? If possible, name how you can test these assumptions. Please justify your answer.*

Figure 3. Example of item 1 in pre-and posttest. Short problem description with three sub-questions to be answered that align with the solution steps of the worked example. No distracting information is given in contrast to the practice problems during the learning phase (translated from German).

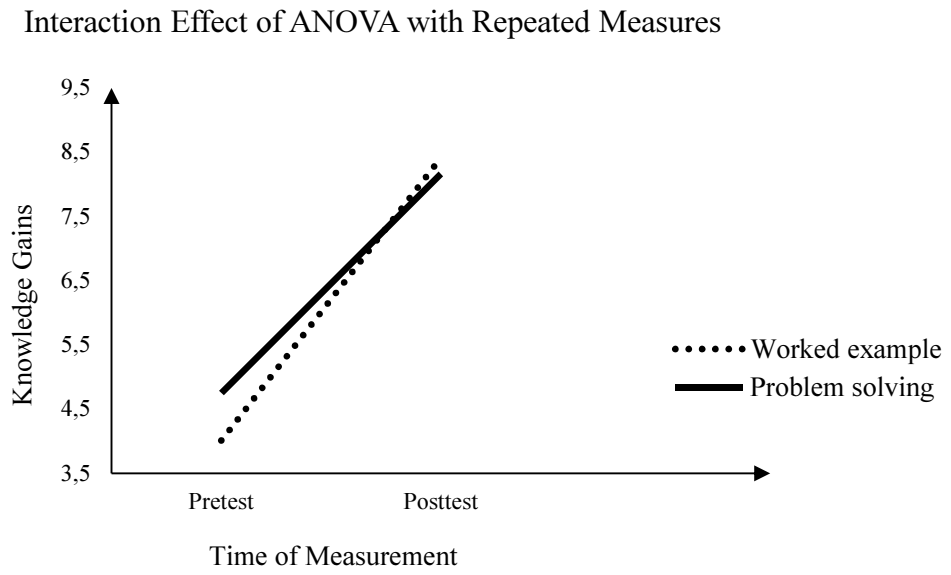

Figure 4. Graphical representation of interaction effect between instructional treatment and time of measurement. All participants improved significantly from pre-to posttest, however, the increase was significantly larger in the condition with worked examples. Thus, worked examples were more effective than problem-solving in this sample.

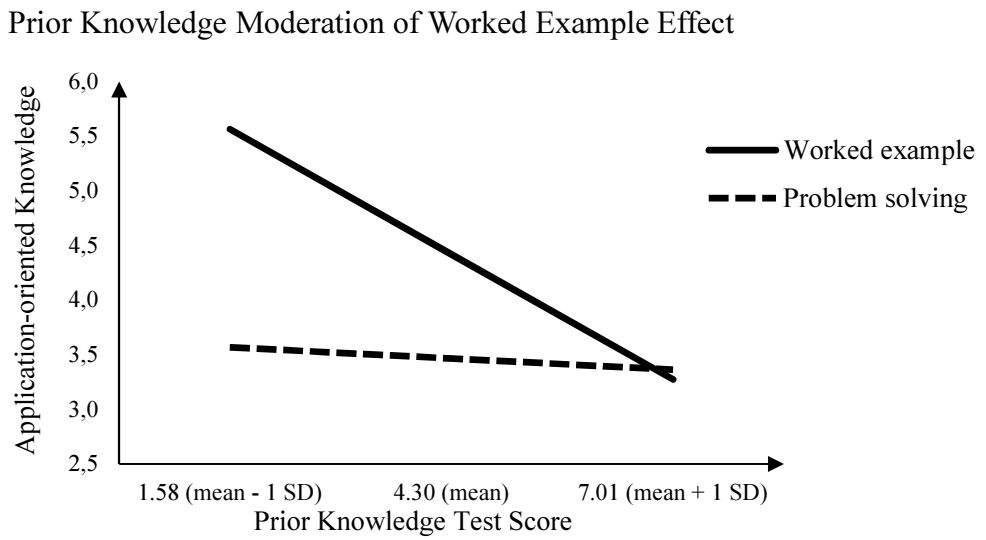

Figure 5. Graphical representation of interaction effect between worked examples and prior knowledge. The effect of worked examples on knowledge gains for three values of the moderator (mean, +/- one SD) are shown to indicate the trend of the moderation.

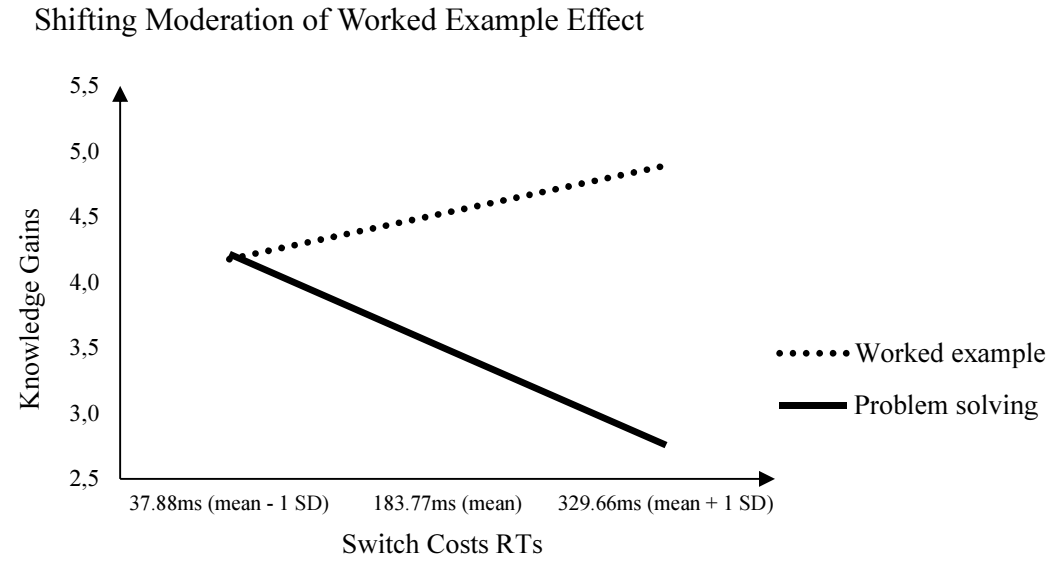

Figure 6. Graphical representation of interaction effect between worked examples and shifting ability. The effect of worked examples on knowledge gains for three values of the moderator (mean, +/- one SD) are shown to indicate the trend of the moderation. Please note that *higher* shifting values indicate *lower* shifting ability.

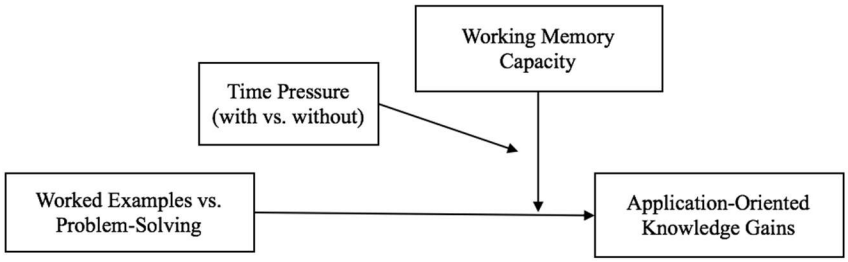

Figure 7. Worked examples vs. problem-solving is the independent, working memory capacity (WMC) the primary, time pressure (with vs. without) the secondary moderator, and application-oriented knowledge gains the dependent variable in this moderated moderation model. Tested was the assumption that WMC only moderates the worked example effect on knowledge acquisition if learners are under time pressure.

Appendix A

Table A1

Example Items for Subscales of the Questionnaire of Current Motivation

Sub-scale	Example Item
Interest	*"I would solve such tasks in my spare time."*
Challenge	*"The task is a real challenge for me."*
Probability of success	*"I think I will not manage to solve this task."*
Anxiety	*"When I think about this task, I get worried."*

Note. Translation from German to English by first author.

Table A2

Questionnaire of Current Motivation Subscale Reliabilities

Subscale	Cronbach's Alpha	Number of Items	N
Interest	.81	5	231
Challenge	.57	4	231
Probability of success	.73	4	231
Anxiety	.86	5	231

Appendix B

Table B1

Descriptive statistics of moderating and dependent variables in the replication condition

Variable	*Mean*	*SD*	*N*
Prior knowledge	4.3[1]	2.72	115
Posttest knowledge	8.29[1]	3.56	115
Knowledge gains	4.0[1]	2.78	115
Working memory capacity	.69[2]	.14	114
Shifting	183.77[3]	145.89	115
Fluid intelligence	.26[4]	.82	115
Cognitive load	6.05[5]	1.22	115

Note. Metrics: [1] Points in the knowledge test. [2] Mean score of proportion of correctly recalled items in operation, reading, and symmetry span. [3] Reaction times in milliseconds. [4] Composite score for fluid intelligence based on the testing software. [5] Subjective rating on Likert-Scale.

Table B2

Descriptive statistics for knowledge and cognitive load measures by experimental condition

	Worked Examples		*Problem-Solving*	
	Mean	*SD*	*Mean*	*SD*
Prior knowledge	4.01	2.39	4.75	3.00
Posttest knowledge	8.41	2.94	8.16	4.11
Knowledge gains	4.40	2.74	3.59	2.78
Cognitive load	6.00	1.27	6.11	1.18

Note. Worked example condition $n_1 = 57$, problem-solving condition $n_2 = 58$; total N = 115 in replication sample. Prior and posttest knowledge min. = 0, max. = 26 points. Cognitive load was indicated on a scale from 0-9 from *very, very low mental effort (1) to very, very high mental effort (9).*