Instructional Science
https://doi.org/10.1007/s11251-022-09579-4

ORIGINAL RESEARCH

Learning to solve ill-defined statistics problems: does self-explanation quality mediate the worked example effect?

Sarah Bichler[1,3] · Matthias Stadler[1] · Markus Bühner[1] · Samuel Greiff[2] · Frank Fischer[1]

Received: 5 November 2019 / Accepted: 3 February 2022

Abstract

Extensive research has established that successful learning from an example is conditional on an important learning activity: self-explanation. Moreover, a model for learning from examples suggests that self-explanation quality mediates effects of examples on learning outcomes (Atkinson et al. in Rev Educ Res 70:181–214, 2000). We investigated self-explanation quality as mediator in a worked examples—problem-solving paradigm. We developed a coding scheme to assess self-explanation quality in the context of ill-defined statistics problems and analyzed self-explanation data of a study by Schwaighofer et al. (J Educ Psychol 108: 982–1000, 2016). Schwaighofer et al. (J Educ Psychol 108: 982–1000, 2016) investigated whether the worked example effect depends on prior knowledge, working memory capacity, shifting ability, and fluid intelligence. In our study, we included these variables to jointly explore mediating and moderating factors when individuals learn with worked examples versus through problem-solving. Seventy-four university students (mean age = 23.83, *SD* = 5.78) completed an open item pretest, self-explained while either studying worked examples or solving problems, and then completed a post-test. We used conditional process analysis to test whether the effect of worked examples on learning gains is mediated by self-explanation quality and whether any effect in the mediation model depends on the suggested moderators. We reproduced the interaction effects reported by Schwaighofer et al. (J Educ Psychol 108: 982–1000, 2016) but did not detect a mediation effect. This might indicate that worked examples are directly effective because they convey a solution strategy, which might be particularly important when learning to solve problems that have no algorithmic solution procedure.

Keywords Self-explanation quality · Worked examples · Executive functions · Moderated mediation · Problem-solving

Matthias Schwaighofer, originally second author of this publication, died on June 10th, 2017. This work would not have been possible without him.

This manuscript is part of the doctoral dissertation of the first author and has partially been presented at conferences prior to manuscript submission. It is based on a study by Schwaighofer, Bühner, and Fischer (2016) and reports data that have been previously published by these authors as well as data that have not been analyzed and published. Data and materials are shared on the Open Science Framework (OSF) https://osf.io/ynr4p/.

Extended author information available on the last page of the article

Introduction

2019 marked 30 years since the seminal work by Chi et al. (1989) was published and set research on the "self-explanation effect" in motion. Ever since, our understanding of this effect has greatly improved. For example, an essential component of self-explaining is that learners make inferences (Rittle-Johnson & Loehr, 2017); learners need support to self-explain in ways that boost learning (Renkl et al., 1998); and when learners are supported, self-explaining is a powerful activity across domains and outcomes such as conceptual or procedural knowledge (Rittle-Johnson et al., 2017). Self-explaining is particularly important in combination with worked examples, whose effect seems to depend on the quality of self-explanations learners generate while studying examples (Chi et al., 1989; Renkl, 1997; Sweller et al., 2019). In fact, in their framework of example-based learning, Atkinson et al. (2000) assume that the effectiveness of worked examples not only depends on self-explanation quality but also that "the structure of worked examples enhances students' self-explanation behavior" and that "students' self-explanation behavior during study in turn mediates learning" (Atkinson et al., 2000, p. 204).

As researchers investigated mostly the effect types of self-explanation prompts have on self-explanation quality or what types of worked examples self-explaining to combine with, we know less about the effect of self-explaining when worked examples are compared to problem-solving. In the worked examples—problem-solving paradigm, it is common to investigate the conditions under which worked examples are more effective than problem-solving. This resulted in robust evidence that worked examples are well suited for learners with less prior knowledge in the domain of study and that problem-solving is well suited for learners with more prior knowledge (Kalyuga, 2007). In this study, we investigate self-explanation quality as mediator in the worked examples—problem-solving paradigm. Drawing on Atkinson et al.'s framework, we assume that learners generate self-explanations of higher quality when studying worked examples compared to when solving problems and because of that gain more knowledge. In detail, we analyzed self-explanations of a study by Schwaighofer et al. (2016).

Schwaighofer et al.'s work focused on moderators of the worked example effect and their study contributed to the field of instructional science by looking not only at prior knowledge but also other cognitive functions (working memory capacity (WMC), shifting ability and fluid intelligence). Our study further contributes to the field as we are applying conditional process analysis (Hayes, 2018), an analytical approach that allows for the joint analysis of mediating and moderating factors. With this study we thus intend to bring forth a discussion of the interdependence of instructional support, cognitive characteristics of the learner who uses the support, and the activity the learner is engaged in while using the support.

Rationale

The self-explanation effect

Ever since Chi et al. (1989) found that learners who frequently referred to domain principles or linked the materials with their prior knowledge during example study were those who did well on a later test and that learners who mostly rephrased the materials

and overall engaged in less self-explaining were those who did less well on the test (Chi et al., 1989), the self-explanation effect refers to the finding that worked examples are more effective when learners self-explain well than when they produce no or low-quality self-explanations (Renkl, 1997). Self-explanations are what learners produce when they explain learning materials to themselves (Chi et al., 1989). Self-explaining is assumed to prevent superficial use of worked examples. It actively involves learners with the learning materials (Roy & Chi, 2005; Stark et al., 2000) and, ideally, learners generate inferences that go beyond what is provided in the instructional materials (Chi, De Leeuw, et al., 1994; Rittle-Johnson et al., 2017), supporting learners to deepen their understanding (Chi & Wylie, 2014). Further, learners attend more to structural than to surface features of problems when self-explaining worked examples. Thus, self-explaining promotes knowledge that is less tied to specific problem features, which helps learners to generalize and to transfer their knowledge to solve problems in new contexts (Adams et al., 2014; Renkl, 2014; van Gog & Rummel, 2010).

Self-explaining is a central component in the framework for learning from examples proposed by Atkinson et al. (2000). The framework describes the effect of worked examples on learning outcomes as dependent on self-explanation quality. This means that examples are only effective if learners self-explain well during example study. The framework also describes that certain features of worked examples support learners to generate self-explanations of higher quality and that engaging in high-quality self-explaining then results in better learning. Thus, self-explanation quality is assumed to mediate, that is "explain" the effect of examples on learning. Based on this framework it seems desirable to design worked examples that elicit high-quality self-explanations or to design supports for students to generate high-quality self-explanations. Focused on the latter, one study compared self-explanation prompts to no prompts (Berthold & Renkl, 2009) and another study compared different types of self-explanation prompts to no prompts (Berthold et al., 2009). Both studies found that the quality of self-explanations accounted for differences between the experimental groups in conceptual and procedural knowledge. Other studies that investigated self-explanation quality as mediator developed training interventions featuring video-based examples to support learners gaining declarative knowledge about argumentation (Hefter, 2021; Hefter et al., 2014) or conceptual knowledge about epistemological understanding and intellectual values (Hefter et al., 2015). Self-explanation quality mediated the effect of interventions on outcomes. However, in these studies, the difference between the experimental and control condition was that the experimental group self-explained the learning domain and the control group the exemplifying domain. For example, the experimental group self-explained how a model used argumentation principles in an argument while the control group self-explained the content about which the model argued (e.g., Hefter, 2021). Further, it may be argued that some types of prompts in Berthold et al. (2009) provided learners with additional information which might have helped learners to better understand the content. Roelle and Renkl (2020) applied proper mediation analysis testing their hypothesis that the effect of access to review of instructional materials versus no access while working on examples on posttest chemistry knowledge was mediated by the number of self-explanations learners generated. The authors included academic self-concept as moderator in the model and reported finding a moderated mediation effect (Roelle & Renkl, 2020). Taken together, the mediation hypothesis formulated in Atkinson et al. (2000) has for a while not been rigorously tested and the field is moving towards doing so by applying state-of-the-art mediation analysis.

Self-explaining in a worked examples—problem-solving paradigm

It seems intriguing to draw on Atkinson et al.'s (2000) framework of example-based learning when comparing worked examples to problem-solving. One reason learners often fail to solve problems is that they do not have practical or applicable knowledge and instead use general knowledge (Ohlsson, 1996). It seems possible that self-explaining problem-solving helps learners to notice when they get stuck and to identify the confusing part of the problem. Further, when self-explaining, learners might realize that a step they took to solve the problem does not continue to make sense (Ohlsson, 1996; Roy & Chi, 2005). This might guide learners to seek more knowledge or help them recognize which knowledge is relevant. In that sense, self-explaining one's own problem-solving might give learners the chance to "repair" their own knowledge (Allwood, 1984; Allwood & Montgomery, 1982; Ohlsson, 1996; VanLehn & Jones, 1993) or integrate information from instructional materials with their own thinking about the problem (Chi, 2000).

Schwonke et al. (2009) investigated self-explanation quality in the worked examples—problem-solving paradigm. They compared faded worked examples to problem-solving in a cognitive tutoring environment. Learners solved geometry problems and thought aloud providing justifications either for the solutions they generated or for the worked-out steps of the example they studied. Analysis of the self-explanations showed that while groups differed on certain types of self-explanations, they did not differ in the total number of self-explanations. However, only the total number of self-explanations correlated with learning outcomes (Schwonke et al., 2009). The mediation effect was not tested. Similarly, a study that compared repeated example study with example-problem pairs, both accompanied by self-explanations (referred to as elaborations), reported how the number of self-explanations as well as different kinds of self-explanations differed between groups and related to the learning outcome (Stark et al., 2000), but did not directly mention the possibility of a mediated effect, nor report a mediation analysis. Hence, directly testing whether self-explanation quality mediates the worked example effect will shed light on how learners who study worked examples outperform learners who solve problems.

Self-explanation quality

Quality of self-explanations has been conceptualized and consequently operationalized very differently across the literature. To distinguish self-explanations from "good" and "poor" learners, Chi et al. (1989) subjected their verbal protocols to an extensive in-depth analysis. Subsequently, rephrased text was used as an indicator of low-quality self-explaining, whereas linking study materials to worked examples, linking materials to one's own prior knowledge, or referencing domain principles were taken as indicators of high-quality self-explaining (Chi et al., 1989; Chi, De Leeuw, et al., 1994; VanLehn & Jones, 1993). Based on this conceptualization, Renkl (1997) defined several categories of self-explanations: For example, principle-based explanations are those in which learners verbalize or refer to a domain principle (physics law, mathematical principle, etc.), whereas goal-operator explanations are those in which learners identify a sub-goal and mention operators that help them to achieve it (Renkl, 1997). These types of self-explanations were used and adapted to the domains or learning tasks in studies that followed (e.g., Berthold & Renkl, 2009; Berthold et al., 2009; Schwonke et al., 2009). A recent study defined principle-based self-explanations as interrelating instructional text and examples, expanding

this type of self-explanation to relating principles and/or concepts from learning materials with concrete examples (Roelle & Renkl, 2020). Often, the quantity of a certain self-explanation type was used as indicator of self-explanation quality (e.g., Berthold & Renkl, 2009; Berthold et al., 2009; Renkl, 1997; Roelle & Renkl, 2020; Schwonke et al., 2009). Others, however, conceptualized quality as the degree to which self-explanations included elements of the to-be-learned materials and rated the correctness of the information in a self-explanation (Hefter et al., 2014, 2015; Schworm & Renkl, 2007). In such approaches, incorrect self-explanations constituted the lower end of the quality continuum. In other studies, incorrect self-explanations were treated as a separate category (e.g., Berthold & Renkl, 2009; VanLehn & Jones, 1993). This means that what constitutes a self-explanation is closely tied to the learning task or material. Consequently, the definition of a high-quality self-explanation greatly varies resulting in a range of operationalizations including quantitative and qualitative measures.

Types of problems and types of knowledge

In prior studies different examples and self-explanations were designed to match the nature of different learning materials and goals. For example, Chi et al. (1989) used science problems (mechanics problems). To solve these, learners had to combine mathematical and deep conceptual knowledge. Many other studies utilized math, probability, or geometry problems that can be classified as procedural problems. To solve these, learners need to apply rule-based or algorithmic procedures (Paas, 1992; Paas & van Merriënboer, 1994; Schwonke et al., 2009). Such problems seem to require understanding rules and principles, as well as the order in which to apply rules. Another type of problems are function-form problems, which require a long causal explanation, for example science problems like photosynthesis or the human circulatory system (Chi, De Leeuw, et al., 1994; Chi, Slotta, et al., 1994). To solve such problems, learners need to apply deep conceptual knowledge. Yet other studies used double content examples to train domain general skills such as collaboration or argumentation. Double content examples operate on two levels. The example needs to model the targeted domain (e.g., argumentation, how to write an essay) which to do requires an exemplifying domain (arbitrary exchangeable topics that are not the learning goal, for example the content a model argues about, the content the essay is on). The aim for learners was to gain knowledge in the target domain (e.g., Hefter et al., 2014; Schworm & Renkl, 2007).

The types of problems that are used and the types of knowledge that are targeted involve learners differently with the learning materials and result in different self-explanations. Mechanics problems such as the ones used by Chi et al. (1989) or procedural problems such as math and probability problems (e.g., Paas, 1992; Paas & van Merriënboer, 1994; Schwonke et al., 2009) may only be solved if certain steps are taken in a specific order. Learners self-explaining worked examples that illustrate the steps for solving such problems need to abstract rules and a stepwise procedure. Thus, learners generate principle-based or goal-operator explanations. In contrast, learners do not need to abstract rules or procedures when learning to solve argumentation problems. Self-explaining what a model does arguing about a certain topic more likely involves noticing of argumentation moves and creating a list of possible moves. Self-explanations tend to refer to domain concepts, their functions, and/or relations when examples serve learners who are trying to gain an understanding of disciplinary systems or phenomena (Chi, De Leeuw, et al., 1994).

The statistics problems Schwaighofer et al. (2016) designed for their study pertain to a fictitious researcher who wants to address a specific question and does not know how to statistically do so. These problems require learners to identify relevant information and apply their knowledge (Schwaighofer et al., 2016). Without an example, a model, or instructions, learners may be unable to infer any rules for how to solve such a problem. Thus, like argumentation problems, these problems do not have an algorithmic solution. Hence, we refer to them as ill-defined. Like function-form problems, the problems designed by Schwaighofer et al. (2016) require conceptual knowledge to be solved. Instead of reflecting on solution steps at a more abstract level, learners reason which statistical concepts can be applied to a given problem, apply their knowledge, and then justify why a certain concept applies. Therefore, self-explanations are likely highly intertwined with the content of the problem solution. Consequently, it is improbable that the self-explanations that occur are of the same type as those that occur when rule-based problems are being solved. For ill-defined statistics problems it is more suited to operationalize quality as the degree to which the solution includes correctly applied and justified conceptual information. This operationalization seems to resemble the recent approach of capturing self-explanations in chemistry learning where students used principles but also concepts from the learning materials and applied them to concrete examples in their self-explanation, even though this was categorized as principle-based self-explanations (Roelle & Renkl, 2020).

The dependency between the type of problem or type of targeted knowledge and self-explanations that is observed raises the question how other features of the learning task determine the quality aspects that can be observed. For instance, learners were asked to self-explain out loud (e.g., Chi et al., 1989), to answer content related questions (e.g., Hefter et al., 2014), to justify their solutions (e.g., Schwonke et al., 2009), or to explain why (Roelle & Renkl, 2020). Are observed differences between the artifacts learners generate in response to "self-explain what you are doing", "explain why","think aloud", or "justify your answer" manifestations of distinct thinking processes or the result of different prompts? Are there quality indicators that can be used across data sources or does the type of data collected (written explanations versus think aloud protocols) determine the type of self-explanation we can observe? The critical feature of self-explaining is its constructive, generative nature (Chi, De Leeuw, et al., 1994). Potentially, self-explaining can be understood as a broad range of activities as long as they include generation of inferences (see Rittle-Johnson et al., 2017). These are important questions in need of further conceptual discussion and empirical investigation.

Moderators of the worked example effect

Our investigation of self-explanation quality as mediator in the worked examples—problem-solving paradigm is grounded in Schwaighofer et al.'s (2016) study. In consensus with the field of worked examples research, Schwaighofer et al. argued that the worked example effect depends on individual learners' prerequisites. Prior knowledge is one such prerequisite (the expertise-reversal effect, Kalyuga, 2007). In a review of worked examples research, van Gog and Rummel (2010) suggested that individual differences in working memory capacity (WMC) similarly moderate the worked example effect. A different view was proposed by cognitive load theorists who argue that individual differences in WMC are only minimally meaningful in complex learning (Paas & Sweller, 2014). Beyond prior knowledge and WMC, other cognitive functions and their impact on instructional effectiveness are rarely discussed (see Schüler et al., 2011). Schwaighofer et al.'s (2016) goal was to empirically investigate WMC as

moderator of the worked example effect and to include multiple cognitive functions relevant to learning as moderators. The authors thus measured prior knowledge, WMC, shifting and fluid intelligence. Their results suggest that shifting ability and fluid intelligence influence the effect of worked examples. Shifting is a core cognitive function needed to regulate thought and behavior (Miyake & Friedman, 2012), as it allows us to flexibly switch between carrying out different tasks (Miyake et al., 2000). Worked examples were more beneficial than problem-solving for learners with lower shifting ability, but the benefit of worked examples over problem-solving decreased with higher shifting ability (Schwaighofer et al., 2016), an effect that was replicated (Bichler et al., 2020). In contrast, no evidence was found for WMC as moderator of the worked example effect (Bichler et al., 2020; Schwaighofer et al., 2016), favoring the cognitive load explanation that everyone's working memory capacity is limited when it comes to complex learning (Paas & Sweller, 2014). While Schwaighofer et al.'s (2016) and the replication of their study (Bichler et al., 2020) extended the investigation of factors that moderate the worked example effect to include prior knowledge and other cognitive functions relevant to learning, neither study analyzed learners' self-explanations or the mediating role of self-explanation quality.

The present study

Motivated by the fact that Atkinson et al.'s (2000) model of learning from examples through self-explanations has not yet been rigorously tested, we use Schwaighofer et al.'s (2016) data to analyze self-explanation quality and test the assumed effect in a proper mediation analysis. Operating within the worked examples—problem-solving paradigm, we do so without ignoring that there are factors that moderate the worked example effect. Specifically, we use conditional process analysis to jointly investigate moderating and mediating factors (Hayes, 2018). Using conditional process analysis, our study contributes to the field of instructional research by delivering results that stimulate rethinking the theoretical assumptions in the worked examples—problem-solving paradigm and around the effect of self-explanations and allow for a contextualized interpretation of the involved moderating and possibly mediating factors.

Further, we synthesized work done in the worked examples—problem-solving paradigm and around the self-explanation effect. We reviewed the types of problems utilized and types of knowledge targeted in previous studies that focused on the self-explanation effect. We contrasted studies using different prompts to elicit self-explanations and analyzing different data sources. Hence, we not only look at self-explanation quality in the worked examples—problem-solving paradigm with a proper mediation analysis, but also provide a report of the field's approaches to conceptualizing and operationalizing self-explanation quality. We model one approach that we argue is suitable for our specific context (ill-defined problems, learners prompted to justify, written artifacts as data source, and application-oriented knowledge as outcome).

Research questions and hypotheses

We address these research questions: (1) Does self-explanation quality mediate the worked example effect on application-oriented knowledge and is the mediation effect dependent on prior knowledge, WMC, shifting ability, and fluid intelligence? and (2) Are the results of

the moderation analysis reported by Schwaighofer et al. (2016) reproduced? We represent the conditional process model that we analyze in Fig. 1.

Our main goal is to test whether self-explanation quality mediates the worked example effect application-oriented knowledge in the domain of statistics. We predict that self-explanation quality mediates the worked example effect on learning gains in such as that self-explanation quality will be higher in the worked example than in the problem-solving group (Fig. 1, path 2) and that higher self-explanation quality will in turn be associated with higher learning gains (Fig. 1, path 3). In other words, we expect an indirect effect (Fig. 1, path 4). Special about conditional process analysis is that the model tests if the indirect effect is contingent on any of the included moderators. In our case these are prior knowledge, WMC, shifting, and fluid intelligence. Such an investigation is novel and explorative; thus, we have no assumptions as to whether the indirect effect would depend on any of the moderators or not (Fig. 1, path 5). Similarly, we have no assumption as to whether the effect of worked examples on self-explanation quality or the effect of self-explanation quality on learning gains would be contingent on any of the moderators (Fig. 1, path 6).

Our secondary goal is to test whether moderation results reported by Schwaighofer et al. (2016) are reproduced in the conditional process model analysis that we conduct. We predict that even when self-explanation quality is included as mediating variable, the same interaction effects that Schwaighofer and colleagues reported are detected. Hence, we predict to find an interaction effect of worked examples versus problem-solving and shifting ability as well as fluid intelligence, and a statistically non-significant interaction of the treatment and prior knowledge as well as WMC (Fig. 1, path 1).

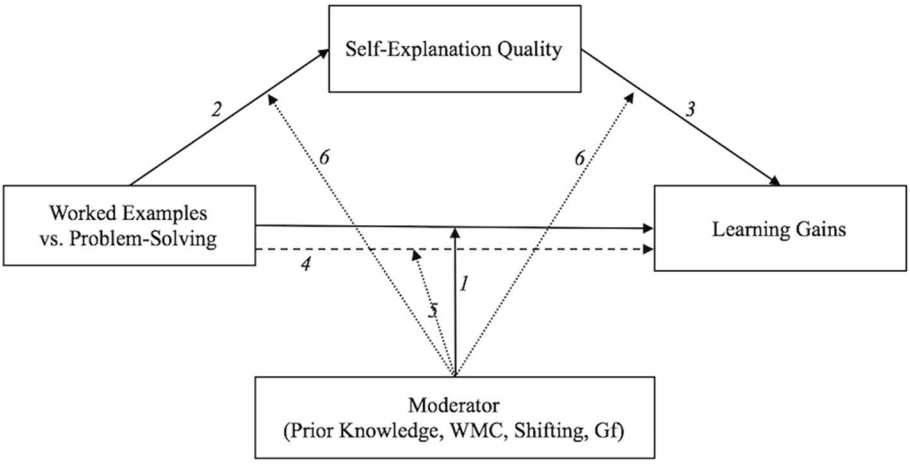

Fig. 1 Conditional process analysis testing whether the worked example effect on learning gains is moderated by prior knowledge, WMC = working memory capacity, shifting ability and Gf = fluid intelligence [1]; whether the worked example effect is mediated by self-explanation quality [4] in such as that worked examples lead to better self-explanation quality [2] and better self-explanation quality is associated with higher learning gains [3]; and testing whether this indirect effect is contingent on any of the four moderator variables [5]; as well as testing whether the a- and b-path is contingent on any of the suggested moderators [6]. The analysis was carried out by testing model 59 in the SPSS Macro PROCESS by A. Hayes (2018). The dashed line represents the indirect effect. Dotted lines are used for those effects that are explored, for all other effects in the model we investigate directional hypotheses

Method

As our investigation pertains to a previously published study, this method section describes the same procedure as is described in Schwaighofer et al. (2016). In "Sample", "Measures (self-explanation quality)", and "Statistical Analysis" we report the unique aspects of our study.

Sample and design

Students enrolled in psychology, educational science, or school psychology undergraduate programs at a large German university participated in Schwaighofer et al.'s (2016) study. Self-explanation data was available for N = 77 participants. One participant had not completed the study and self-explanation worksheets of 2 participants could not be matched with the rest of the data. Thus, the final sample size for the conditional process analysis is N = 74 (88% women, 12% men; mean age = 23.83 years (*SD* = 5.78)). The semester ranged from the second to eighth semester; 29 participants were enrolled in educational science, 24 in psychology, and 21 in school psychology. Participants either received monetary compensation or a participation certificate, which is required to receive study credits in psychology. Participants were randomly assigned to one of two experimental conditions: worked examples ($n_1 = 36$) versus problem-solving ($n_2 = 38$). Participants studied in a lab with a researcher present and were thus observed working on the task, no participant was observed not engaging with the learning materials (those who did not write a lot, studied the slides on the screen or scrolled between the information materials, which was seen as an indicator of engagement).

Material

Material was presented on a computer screen as PowerPoint slides. The first slide of each problem contained a problem description and the research question of a fictitious researcher. Participants were asked: "How can you investigate this problem statistically? Justify your answer, if possible." To solve the problem, participants had to identify the relevant variables as independent, dependent, or control variables; suggest an appropriate design and statistical analysis; and consider relevant statistical assumptions. The problem slide was followed by two slides with textbook material on the General Linear Model, containing information relevant and information irrelevant for the respective problem. Learners had to reason which information applied, thus the task can be described as complex. In total, there were three problems and textbook material for each.

Worked examples

In the problem-solving condition, participants solely had the textbook material to rely on for solving the problems. In the worked example condition, a worked example was additionally provided for each problem. Each worked example consisted of three consecutive slides inserted between the problem description and the textbook material. The first slide of the worked example included the first solution step: "Identifying variables and design". This was the heading of the slide. The body of the slide showed the solution: the

independent, dependent, control variables, and design of the respective problem. The second slide showed the second solution step: "Choosing a statistical method" and its application to the problem; the third slide showed the third solution step: "Relevant statistical assumptions" and the solution for this step.

Procedure

The experiment was split into two sessions, amounting to a total of 4 h of testing. In the first session, participants were introduced to the procedure of the study, provided demographic information, completed the prior knowledge test (paper–pencil), computerized working memory and shifting tasks, as well as the fluid intelligence test battery. The first session took about 160 min. Personal identification codes were used in every task to ensure anonymity and correct matching of data from different sources. The second session consisted of the learning phase and post-test (paper–pencil, 20 min). During the 60-min learning phase, participants were randomly allocated to one of the two conditions and solved the three statistical problems. Participants were able to navigate through all slides without any restrictions but were asked to only move on to the next problem once they had finished the previous one.

Measures

Self-explanation quality

In both conditions, participants responded to this prompt for each problem: "How can you investigate this problem statistically? Justify your answer, if possible." We classify these written artifacts (problem solutions including justifications) as self-explanations. To capture learners reasoning, we used expert solutions to the problems and identified the claims that could be made per problem, as well as the information that would justify each claim. For each problem, three claims could be made, and each claim corresponded to one of the three modeled solution steps in the worked example: Identify (1) variables and design, (2) statistical analysis, and (3) statistical assumptions. For example, a learner who suggested a 2×2-factorial between-subjects design identified an appropriate operator to complete step two of the problem solution. We coded each claim with either no claim (0: no claim or an incorrect claim), incomplete claim (0.5: partially correct or missing details), or complete claim (1: correct and complete). For example, one correct claim for "Statistical Analysis" in one problem was a 2×2-factorial between-subjects design. If a participant suggested a between-subjects design, it was coded with 0.5 because details were missing. If a participant suggested a within-subjects design, it was coded with 0 for incorrect.

We separately coded whether each claim was justified. A claim was justified if information that corresponded to a claim was either visually, spatially, or verbally linked with the claim. For example, a participant used arrows pointing from the variables they identified to the statistical analysis they suggested (visual link). If a participant suggested a statistical analysis and listed the variables below using bullet points, we considered information spatially linked. Claims were verbally linked to justifications if participants used words like "consequently" or "therefore". We coded "justified" if claims were clearly linked to corresponding information (1). If justifying information was not linked to claims or incomplete, we coded "incomplete justification" (0.5) and if learners did not provide any justifying information, we coded 0 for "not justified".

Next, we transformed the separate scores for claim and justification as shown in Table 1 to capture the degree to which each single claim was justified. The logic of this transformation is that a learner who provided no claim and no reason scored lowest (0) and a learner who provided a complete claim that was fully justified scored the highest (3). Using this scoring system, learners could get a minimum of 0 and a maximum of 9 points per problem. A 0 would indicate that this learner did not provide any or any correct claim and did not or incorrectly justify it. A 9 indicates that a learner made 3 claims and fully justified them. We used the sum score across all three problems (all 9 claims) as indicator of self-explanation quality (with three problems a minimum of 0 and a maximum of 27 points was possible). Cronbach's $\alpha = .78$ indicates a good reliability for the 9-item self-explanation quality measure.

We additionally coded when learners applied conceptual information, for example "independent variable", to specific problem information, for example "motivation". We separated this from self-explanation quality and refer to it as *concept application*. The number of relevant concepts varied between the three problems (nine concepts in problem 1 and 3; seven concepts in problem 2). We coded whether a concept was applied to problem information (1 point) or not present (0 points). We gave 0.5 points if either only the conceptual or the problem information was present. For example, "independent variable = motivation" was coded with 1 point. Self-explanations that included only "there is one independent variable" and did not specify that the independent variable was motivation received 0.5 points. We calculated the sum across all concept application points across all three problems. A minimum of 0 and a maximum of 25 points could be achieved. Cronbach's $\alpha = .82$ indicated a good reliability for this scale.

Application-oriented knowledge

Application-oriented knowledge was assessed at pretest and post-test with two open-ended questions that were chosen from an exam question pool of the psychology department. Both questions had to be answered by completing three sub-tasks that aligned with the

Table 1 Self-explanation quality scoring

Claim		Reason	Justification Score
Complete (1)	+	Complete (1)	3
Incomplete (0.5)	+	Complete (1)	2.5
Complete (1)	+	Incomplete (0.5)	2
Incomplete (0.5)	+	Incomplete (0.5)	1.5
Complete (1)	+	No reason (0)	1
Incomplete (0.5)	+	No reason (0)	0.5
No claim (0)		–	0

Note. Each problem had three claims, thus per problem a max. of 9 points could be achieved. The self-explanation quality is defined as the total sum score across 9 possible claims across the 3 problems, with a min. of 0 and a max. of 27 points. Example: A learner mentions Claim 1 correctly and fully justifies it = 3; Claim 2 is incomplete and not justified = 0.5; Claim 3 is complete but not justified = 1. This learner has a final self-explanation quality score of 3 + 0.5 + 1 = 4.5 across the three possible claims and their justifications

three solution steps in the worked examples: (1) identifying variables and design, (2) statistical analysis, and (3) statistical assumptions. In total, participants could receive 6 points (one for each sub-task in both test questions). Items in pre- and post-test were structurally the same, but the content of the items changed from pre- to post-test. For example, the number of independent variables and their scale level was the same in pre- and post-test but in the pretest the independent variable was motivation and in the post-test it was intelligence.

As reported by Schwaighofer et al. (2016), the pre- and post-test were coded with a good interrater reliability (average Cohen's $\kappa = .95$ and $\kappa = .92$ respectively), however the Kuder-Richardson-20 coefficient indicated a relatively low reliability for the pretest $r_{tt} = .49$ and the post-test $r_{tt} = .44$ (N = 76). The difference between test scores (post-test—pretest) was used as the dependent variable in the analyses reported by Schwaighofer et al. (2016). We used the same gain scores as the dependent variable in our analyses if not indicated otherwise (different N due to the smaller sample for which self-explanation data was available).

Executive functions and fluid intelligence

Schwaighofer et al. (2016) assessed working memory capacity (WMC) with the automated operation span, the automated reading span, and the automated symmetry span task (Redick et al., 2012) using E-Prime software (version 2.08.22). Participants must, for example, recall the order of letters presented after a processing activity. For each task the proportion of correctly recalled elements was computed as WMC indicator. The mean across these three indicators was then used as measure of WMC. Shifting was assessed with the color-shape task, number-letter task, and category-switch task (Friedman et al., 2011) using the procedure of Friedman et al. (2016). These tasks comprise of switch (applying a rule different to the previous rule) and no-switch trials (applying the same rule in consecutive trials). The mean reaction time of no-switch trials was subtracted from the mean reaction time of switch trials to estimate switch costs for each of the three tasks. A mean score across the three tasks was then calculated as the shifting measure (higher scores indicating lower ability). Fluid intelligence was assessed with three subtests of the computerized intelligence structure battery (INSBAT; Arendasy et al., 2012). The test was adaptive, adjusting the number and difficulty of tasks to the level of performance for each participant. A raw score for fluid intelligence was automatically calculated by the test system, higher scores indicating higher fluid intelligence. Schwaighofer et al. (2016) provide a more detailed description of these measures.

Statistical analyses

We report results based on a 5% alpha level in this paper. Correlations and descriptive statistics were obtained with SPSS 24 (IBM SPSS Statistics, Version 24). Conditional process analyses, also referred to as moderated mediation analyses, were estimated with the SPSS macro PROCESS (Hayes, 2018). We report unstandardized regression coefficients in conditional process analyses and 95% bootstrapped (5000 samples) confidence intervals (CI_{boot}). We report *p*-values for one-tailed testing for the effects that Schwaighofer et al. (2016) formulated as directional hypotheses, as well as for the treatment effect on the mediator and the mediator effect on the outcome. For these effects, we estimated the 90% two-tailed CI_{boot} as an approximation for the one-tailed 95% CI_{boot}. While we assumed a directional mediation effect, we had no assumptions as to whether the indirect effect would

depend on the suggested moderators and thus report results for two-tailed tests. Likewise, two-tailed tests are reported for those effects in the model that we explored (interaction of treatment and suggested moderators predicting self-explanation quality as well as interaction of self-explanation quality and suggested moderators predicting the outcome). The N in all reported analyses is 74 unless indicated otherwise. Schwaighofer et al. (2016) identified age and semester as covariates for their analyses, which is why we also controlled for these variables in our analyses. Conditional process analyses include analysis of new (self-explanation) and previously published data (application-oriented knowledge, moderators). Descriptive statistics and correlations are only reported for self-explanation data (new data). Histograms and boxplots showed that self-explanation quality was approximately normally distributed and that there were no outliers.

Results

Correlations

Self-explanation quality and concept application strongly correlated $r = .72$, $p < .001$. We continue reporting results for self-explanation quality as it is the mediating variable of interest in this study. Self-explanation quality correlated significantly with prior knowledge $r = .24$, $p = .042$ and post-test knowledge $r = .37$, $p = .001$, but not with gain scores $r = .16$, $p = .188$. None of the correlations with covariates and moderating variables (except prior knowledge) were statistically significant (see Table 2). Detailed descriptive statistics for the process measures are found in Table 3.

Conditional process analysis with prior knowledge as moderator

We investigated whether the worked example effect was mediated via self-explanation quality and whether the indirect effect or any other effect in the model depended on prior knowledge. This moderated mediation model included worked examples versus problem-solving as independent variable, self-explanation quality as mediator, post-test knowledge as dependent variable, and prior knowledge as moderator. Age and semester were included as control variables.

The worked example effect on self-explanation quality was not statistically significant $b = 1.78$, $p_{one\text{-}tailed} = .234$, 95% CI_{boot} [–2.58, 5.38]. Similarly, the mediator self-explanation quality did not significantly predict post-test knowledge $b = 0.07$, $p_{one\text{-}tailed} = .059$, 95% CI_{boot} [–0.01, 0.12]. The direct effect of worked examples on post-test knowledge

Table 2 Correlations of self-explanation quality (mediator) with control and moderating variables

		Control variables		Moderating variables			
		Age	Semester	Prior knowledge	WMC	Shifting	Gf
Quality of self-explanations	*r*	– .22	.11	.24	.11	.05	.15
	p	.060	.354	.042	.346	.657	.217

Note that higher scores for shifting indicate lower ability

WMC working memory capacity, *Gf* fluid intelligence

Table 3 Descriptive statistics for sample and per condition

Self-explanation quality	Full sample			
	M	SD	Min	Max
	12.59	6.19	0	24.50
	Worked examples		Problem-solving	
	M	*SD*	*M*	*SD*
	13.56	1.05	11.67	0.97
Concept application	Full sample			
	M	SD	Min	Max
	11.78	5.0	1	20.50
	Worked examples		Problem-solving	
	M	SD	M	SD
	11.74	0.84	11.83	0.80

Note. Descriptive statistics for the process variables concept application and self-explanation quality (mediator in subsequent analyses) for full sample and per condition. Descriptively, worked examples helped learners (on average) to improve their self-explanation quality, while learners in both conditions were on average similarly able to apply statistical concepts

was not statistically significant $b = 0.65$, $p_{one\text{-}tailed} = .120$, 95% CI_{boot} [–0.17, 1.54]. The direct effect of worked examples on post-test knowledge was not moderated by prior knowledge $b = -0.02$, $p_{one\text{-}tailed} = .473$, 95% CI_{boot} [−0.62, 0.43]. Likewise, no interaction effect of prior knowledge and worked examples on self-explanation quality $b = 0.35$, $p = .742$, 95% CI_{boot} [−.41, 2.88] or of prior knowledge and self-explanation quality on post-test knowledge $b = -0.01$, $p = .794$, 95% CI_{boot} [–0.04, 0.06] was found. The indirect effect of worked examples through self-explanation quality was not statistically significant in this model (mean –1 SD: $b = 0.13$, $SE_{boot} = 0.165$, 95% CI_{boot} [−.17, 0.48]; mean: $b = 0.14$, $SE_{boot} = .117$, 95% CI_{boot} [−.03, 0.42]; mean + 1 SD: $b = 0.15$, $SE_{boot} = 0.182$, 95% CI_{boot} [−0.05, 0.63]).

In sum, we found no evidence for an indirect effect of worked examples through self-explanation quality or its dependence on the moderator prior knowledge. Likewise, the direct effect of worked examples on post-test knowledge was not moderated by prior knowledge in this study.

Conditional process analysis with WMC as moderator

To investigate whether the worked example effect was mediated via self-explanation quality and whether the indirect effect or any other effect in the model depended on WMC, we included in our next moderated mediation analysis: worked examples versus problem-solving as independent variable, self-explanation quality as mediator, learning gains (post-test minus pretest scores) as dependent variable, and WMC as moderator. We controlled for age, semester, and fluid intelligence.

In this model, the worked example effect on self-explanation quality was not statistically significant $b = -4.98$, $p_{one\text{-}tailed} = 0.256$, 95% CI_{boot} [−6.58, 5.62] and self-explanation quality did not significantly predict learning gains $b = 0.17$, $p_{one\text{-}tailed} = .035$, 95% CI_{boot} [–0.03, 0.33]. The direct effect of worked examples on learning gains was not statistically significant $b = 1.00$, $p_{one\text{-}tailed} = .224$, 95% CI_{boot} [–1.19, 3.07] and not moderated by WMC $b = -0.35$, $p_{one\text{-}tailed} = .428$, 95% CI_{boot} [−3.54, 2.77]. Similarly, the interaction effect

of WMC and worked examples on self-explanation quality $b = 10.57$, $p = .316$, 95% CI_{boot} [−6.66, 28.80] and of WMC and self-explanation quality on learning gains $b = -0.18$, $p = .182$, 95% CI_{boot} [−0.45, 0.07] were not statistically significant. The indirect effect of worked examples through self-explanation quality was not statistically significant at different levels of the moderator WMC (mean –1 SD: $b = 0.04$, $SE_{boot} = .177$, 95% CI_{boot} [−0.34, 0.40]; mean: $b = 0.10$, $SE_{boot} = .104$, 95% CI_{boot} [−0.05, 0.35]; mean + 1 SD: $b = 0.04$, $SE_{boot} = .192$, 95% CI_{boot} [−0.28, 0.51]).

In sum, we found no evidence for an indirect effect of worked examples through self-explanation quality or its dependence on the moderator WMC in this model. Further, the direct effect of worked examples was not dependent on WMC.

Conditional process analysis with shifting as moderator

The conditional process analysis that investigated the indirect effect of worked examples through self-explanation quality and shifting ability as moderator included worked examples versus problem-solving as independent variable, self-explanation quality as mediator, learning gains as dependent variable, and shifting ability as moderator, as well as age and semester as control variables.

Worked examples did not significantly predict self-explanation quality $b = 0.71$, $p_{one\text{-}tailed} = .408$, 95% CI_{boot} [−4.14, 5.47] and self-explanation quality did not predict learning gains $b = 0.06$, $p_{one\text{-}tailed} = .102$, 95% CI_{boot} [−0.01, 0.13]. Similarly, we found no direct effect of worked examples on learning gains $b = -0.41$, $p_{one\text{-}tailed} = .212$, 95% CI_{boot} [−1.24, 0.48]. However, the direct effect was moderated by shifting ability $b = 0.005$, $p_{one\text{-}tailed} = .002$, 95% CI_{boot} [0.001, 0.01]. The explored interaction between worked examples and shifting ability in predicting self-explanation quality $b = 0.01$, $p = .482$, 95% CI_{boot} [−0.01, 0.02], as well as the interaction between self-explanation quality and shifting ability in predicting learning gains $b \leq 0.001$, $p = .429$, 95% CI_{boot} [−0.001, 0.0002], were both statistically non-significant. Like in the previous models, we detected no indirect effect or its contingency in the model with shifting ability as moderator (mean –1 SD: $b = 0.06$, $SE_{boot} = .131$, 95% CI_{boot} [−0.16, 0.39]; mean: $b = 0.06$, $SE_{boot} = .092$, 95% CI_{boot} [−0.05, 0.31]; mean + 1 SD: $b = .02$, $SE_{boot} = .182$, 95% CI_{boot} [−0.23, 0.52]).

Thus, while there was no evidence for an indirect effect of worked examples through self-explanation quality in this model, the assumed moderation of the direct effect was reproduced (Schwaighofer et al., 2016). As shifting ability is scored such that higher scores reflect lower shifting ability, the interaction effect indicates that the benefit of worked examples over problem-solving increases with decreasing shifting ability. Thus, worked examples seem to be particularly beneficial for learners with lower shifting ability.

Conditional process analysis with fluid intelligence as moderator

We estimated a moderated mediation model with worked examples versus problem-solving as independent, self-explanation quality as mediating, learning gains as dependent, and fluid intelligence as moderating variable, and controlled for age, semester, and WMC.

Worked examples were not found to be predictive of self-explanation quality $b = 1.42$, $p_{one\text{-}tailed} = .190$, 95% CI_{boot} [−1.06, 4.04], and self-explanation quality was not predictive of learning gains $b = 0.04$, $p_{one\text{-}tailed} = .070$, 95% CI_{boot} [−0.01, 0.09]. In this model, there was a statistically significant difference between worked examples and problem-solving with respect to learning gains $b = 1.14$, $p_{one\text{-}tailed} = .002$, 95% CI_{boot} [0.52, 1.74]. This effect

was moderated by fluid intelligence $b = -0.92$, $p_{one\text{-}tailed} = .018$, 95% CI_{boot} [−1.61, − 0.27]. With respect to the explorative part of the model, fluid intelligence did not interact with worked examples predicting self-explanation quality $b = 1.82$, $p = .272$, 95% CI_{boot} [−1.77, 4.75] and likewise did not interact with self-explanation quality predicting learning gains $b = 0.01$, $p = .804$, 95% CI_{boot} [–0.05, 0.08]. The indirect effect of worked examples through self-explanation quality was not contingent on fluid intelligence (mean −1 SD: $b = 0.03$, $SE_{boot} = .107$, 95% CI_{boot} [−0.14, 0.31]; mean: $b = 0.11$, $SE_{boot} = .109$, 95% CI_{boot} [−0.04, 0.38]; mean + 1 SD: $b = 0.21$, $SE_{boot} = .239$, 95% CI_{boot} [−0.08, 0.83]).

Taken together, a difference in learning gains between worked examples and problem-solving was detected, but self-explanation quality did not predict learning gains, and no indirect effect of worked examples through self-explanation quality was found. As expected, the direct effect of worked examples was contingent on fluid intelligence. Specifically, the difference between worked examples and problem-solving in learning gains decreased with increasing fluid intelligence. Thus, worked examples seem to also be particularly beneficial for learners with lower fluid intelligence. Hence, the moderation effect reported by Schwaighofer et al. (2016) was reproduced.

Discussion

We analyzed self-explanation data from a study conducted by Schwaighofer et al. (2016) to investigate self-explanation quality as mediator in the worked examples—problem-solving paradigm. We estimated conditional process models (Hayes, 2018) to jointly investigate moderating and mediating factors of the worked example effect. We assumed that the worked example effect on learning gains would be mediated by self-explanation quality. Beyond our expectations that the moderation effects reported by Schwaighofer et al. (2016) would be reproduced, we had no specific assumptions as to which other effects in the model were dependent on any of the suggested moderators (prior knowledge, WMC, shifting ability, and fluid intelligence). Overall, the developed measure of self-explanation quality was reliable. We measured whether students justified the claims they made and separately coded when students applied conceptual to problem information. Solely coding concept application would have given high scores to students who copied the entire worked example to their explanation. We separated self-explanation quality from concept application to make sure to capture the inferences learners made.

The moderation effects reported in Schwaighofer et al. (2016) were reproduced. We found that worked examples are particularly beneficial for learners with lower shifting ability and for learners with lower fluid intelligence while there was no evidence for prior knowledge and WMC moderating the worked example effect. With respect to our main question, we did not find evidence for the indirect effect of worked examples through self-explanation quality. Further, none of the explored effects in the conditional process analyses were dependent on the suggested moderators.

Self-explanation quality as mediator

Inspired by Atkinson et al.'s (2000) model of learning from examples in which effects of examples on learning outcomes are described to be mediated by self-explanation quality, we asked whether self-explanation quality would mediate effects on learning outcomes in a worked examples—problem-solving paradigm. Specifically, we assumed that learners who

self-explain a worked example would generate self-explanations of higher quality in comparison to learners who self-explain their problem-solving, which in turn would account for differences in learning gains between the two groups. In contrast to our assumption, we did not detect an indirect effect of worked examples through self-explanation quality. We found that learners in both groups generated self-explanations of similar quality. Further, self-explanation quality was not predictive of learning. These findings partially align with those of previous studies. Schwonke et al. (2009) did not detect a difference in self-explanation quality, operationalized as the total number of self-explanations provided, between example study and problem-solving participants. However, the total number of self-explanations was positively correlated with outcome measures (Schwonke et al., 2009). Another study found differences between a worked example group and an example-problem pair group, but only on one type of self-explanation, and self-explanations in general were not related to the learning outcome (Stark et al., 2000). Given that these studies did not directly test a mediation effect and that we did not find evidence for an indirect effect of worked examples through self-explanation quality, we are considering that self-explanation quality is not the factor explaining benefits of worked examples over problem-solving. However, the effect of self-explanation prompts was mediated by self-explanation quality on procedural, conceptual or declarative knowledge in other studies (Berthold & Renkl, 2009; Berthold et al., 2009; Hefter, 2021; Hefter et al., 2014, 2015). This difference may stem from the fact that learners in the control group used time and effort to self-explain the exemplifying domain while learners in the experimental groups used time and effort to self-explain the learning domain. As self-explanation quality was operationalized as "…containing only correct self-explanations including all relevant aspects (i.e., argumentative elements and their functions)…" (Hefter et al., 2014, p. 940) it seems plausible that the experimental group generated much more high quality self-explanations. In our study, all learners self-explained the target material. Hence, our study implemented a more rigorous comparison relative to studies in which learners in the control condition did not self-explain in the learning domain. Further, the type of knowledge targeted differs between these and our study. For example, Hefter et al., (2014) found an indirect effect of their intervention via self-explanation quality on declarative knowledge. Learners who were more successful in recognizing and naming elements of an argument in their self-explanation were more successful in recalling these elements in a delayed post-test. Self-explanation quality in our study was operationalized as justified claims and we assessed application-oriented knowledge. Hence, whether self-explanation quality is a mediator of the worked example effect may depend on how quality is conceptualized and what the targeted outcome is. The discrepancy to other studies that found indirect effects might also be explained by the different paradigm in which they were conducted. When all learners study examples or when they learn with different kinds of examples, self-explanation quality might mediate the effect as suggested by Atkinson et al. (2000). For example, Roelle and Renkl (2020) tested the effect of examples with versus without review of instructional materials on chemistry knowledge and included self-explanations as mediating and self-concept as moderating variable. The authors interpret their results as a moderated mediation.

Making sense of self-explanation quality as mediator in the worked examples—problem-solving paradigm, we can use what is known about learning from examples to explain our results. In general, worked examples support learners to abstract rules (Sweller, 1988) and to understand domain principles, which in turn contributes to schema formation (Renkl, 2014) —if learners actively process the examples (Chi, 2000). As learners self-explained in our study, we assume that learners actively processed the examples. Yet, all learners self-explained in our study. Problem-solving becomes more effective when learners self-explain

(Rittle-Johnson, 2006). Hence, we assume that both groups in our study actively processed the learning materials, which would explain why both groups performed equally well in justifying their solutions during the learning phase. We also think that generating self-explanations supported learners in both groups to identify knowledge gaps, another benefit of self-explaining (Roy & Chi, 2005; VanLehn & Jones, 1993). Thus, we perceive the main difference between our conditions that only the worked example group was supported to abstract a solution procedure. Given the ill-defined nature of our problems, abstracting a solution schema might be particularly important. This might explain why both groups self-explained comparably well, but also why the worked example group outperformed the problem-solving group with respect to the learning outcome. Alternatively, learners in the worked example group may have applied analogical reasoning to solve the post-test problems— "…mapping from the example solution to the current problem as directly as possible" (Reimann & Schult, 1996, p. 127) for which learners in the problem-solving group were less well prepared.

In sum, our study highlights that the mediating effect of self-explanation quality is yet to be investigated further to learn more about why learners supported by worked examples outperform learners who problem-solve. With the evidence at hand, we conclude that it is more important for learners to abstract a solution procedure or "see" one way the problems can be solved than to generate high quality self-explanations. Learners in our problem-solving group generated justified claims and solved practice problems like learners in the worked example group, however, our problem-solvers may have solved the problems rather intuitively and did not recognize the parallels in the solution to each problem.

Assessing self-explanation quality

Our review and results raised the question as to whether the kind of data that is collected to assess quality of self-explanations determines the kind of quality aspects that can be observed. For instance, if worked examples omit some information or reasoning, a learner's self-explanation might focus on that gap in the example. Hence, self-explanations fill in missing information (Chi et al., 1989; Roy & Chi, 2005). If, however, worked examples model the entire solution procedure and provide the problem solution, it is plausible that learners do not have to generate as much on their own. In such cases, learners are still required to make inferences, which is considered the defining process of self-explaining (Rittle-Johnson et al., 2017). For example, when trying to understand the circulatory system, learners have to understand single information elements, but also their functions and relations (Chi, De Leeuw, et al., 1994). Similarly, learners in our study had to understand domain concepts such as what is an independent variable, but also once understood, had to apply it to the given problem (recognize the independent variable in a complex problem description). This involves making inferences as to whether or not or why a certain concept applies to a specific case. Further, by providing justifications for their solutions, learners had to add to the materials as they needed to think about how specific concepts relate to each other (Chi, De Leeuw, et al., 1994). The self-explanations generated in these different contexts differ and thus invite different scoring approaches. The conceptual and operational differences in the assessment of self-explanation quality are intriguing and a structured synthesis including a discussion of the generalizability of findings from self-explanation effect research seems a worthwhile future investigation.

Just like the prompt used, the type of problem used, or the kind of data collected calls for different conceptualization and hence operationalization of self-explanation quality,

scoring approaches also seemed to depend on the learning goal (e.g., a procedural skill or knowledge about a topic), the domain (e.g., science, math, argumentation, etc.), or the type of knowledge targeted (declarative vs application-oriented knowledge, for example). For instance, goal-operator explanations are seen as indicators of good self-explaining when mathematics or physics laws help to solve problems (e.g., Chi et al., 1989; Renkl, 1997). If the goal of a study is to foster conceptual understanding, high-quality self-explanations are those that refer to relations (Chi, De Leeuw, et al., 1994), include relevant and correct aspects of the learning domain (Hefter et al., 2014), or include justified claims. Our study highlights that learners need to apply conceptual information and make inferences about how specific information relates to solve ill-defined problems in the domain of statistics. We developed a rigorous coding scheme and opted not to score if relevant information was correctly listed in a self-explanation but rather to score if learners justified claims they made in their solution. We suggest that this approach is useful for assessing self-explanation quality in nonprocedural domains or for problems without algorithmic solution procedures.

Jointly investigating mediating and moderating factors

We reproduced the moderation effects reported by Schwaighofer et al. (2016) in our conditional process models. Thus, even when self-explanation quality is included in the model, the direct effect of worked examples on learning gains is dependent on shifting ability and fluid intelligence, but not on WMC and prior knowledge. These moderation effects are discussed in detail in Schwaighofer et al. (2016) and a replication of their study (Bichler et al., 2020).

We found that cognitive characteristics of a learner mostly influence the effectiveness of instructional support. If learning activities that play a mediating role are investigated separately from these cognitive characteristics, studies might fail to learn about how cognitive characteristics and learning activities interact. As we want our results to generalize to authentic learning situations, it is crucial to model the complexity of real-life learning situations as best as possible. Conditional process analysis seems to be a suitable methodological approach to model this complexity to some degree. At least, if such methods are available, the field should make more use of them. As we know that the worked example effect depends on certain factors, ignoring these while investigating the effect of self-explanations means risking to produce distorted results.

Limitations

To estimate post-hoc power for the indirect effect investigated in this study, we estimated a simple mediation model (worked examples vs. problem-solving as the independent variable, self-explanation quality as mediator, and learning gains as the dependent variable; the latter two z-standardized). Using the obtained standardized model coefficients, $N = 74$, 5000 replications with 20,000 Monte Carlo draws per replication, and a 95% CI, power for the indirect was estimated to be $\approx 30\%$ (Shiny App run through the software package R, developed by Schoemann et al., 2017). As these results were obtained post-hoc and under the premise that the values obtained in our study represent the true effect, they must be interpreted with caution. However, we must consider that the null finding might be due to a power issue. Our sample size is typical for studies in the field of instructional science, or

at least comparable to (Berthold & Renkl, 2009; Berthold et al., 2009; Hefter et al., 2014, 2015) or bigger than (Roelle & Renkl, 2020) sample sizes in studies that detected indirect effects. Hence, this power simulation emphasizes the need for discussing power and moving towards larger studies to reduce the risk of false positive results and increase confidence in evidence for the effects the field is interested in. Discussion is specifically needed to come up with effective strategies as a priori power analyses for complex designs such as moderation and mediation models are anything but trivial (Aguinis, 1995; Aguinis et al., 2005; Lakens & Evers, 2014; Preacher & Sterba, 2019; Schoemann et al., 2017).

The pretest of application-oriented knowledge did not differentiate well between our participants and thus the low reliability of our measure might have contributed to the null finding in our moderation analysis including prior knowledge. We suggest using more items with a broader range of difficulty in future research to increase the variability in prior knowledge. However, it is also noteworthy that internal consistency is a good indicator of reliability when the underlying construct is assumed to be unidimensional, which in case of testing knowledge or understanding must not always, or is often not the case (Taber, 2018). The test items used in this study were taken from the question pool for the statistics exam at the psychology department the authors are affiliated with; because they have been developed by experts and revised multiple times, the items can be assumed to be valid and authentic.

We have included factors as moderators in our study that are known to vary between individuals. It could be argued that variability in prior knowledge is conceptually different from variability in fluid intelligence. Our results may be viewed as limited as we have generalized across these variables and treated them equally in our analytic approach. If and how such conceptual differences could be statistically modelled is a question for future research. Further, we hypothesized linear effects between worked examples, problem-solving, self-explanation quality, the moderators, and the outcome. Arguably, future research should explore non-linear relationships such as whether there are tipping points at which either self-explaining becomes effective or ineffective or at which cognitive functions matter or do not matter anymore. Such future work may also address whether multi-causation is at play by modeling the instructional conditions, self-explanation quality, and all four moderator variables together. Additionally, with sampling a population of university students we may have been able to capture variability in basic cognitive functions only to a certain degree. Hence, our results should conservatively be interpreted for the sampled population only and not be generalized to, for example, school students.

Future research

While future research on instructional effects will benefit from applying statistical methods to test theoretical assumptions that have not yet undergone rigorous empirical tests in general, it seems specifically interesting to follow up on whether worked examples have a direct effect on learning as suggested in our study. Thus, future research should investigate whether it is the abstraction of a solution procedure in the case of problems with no algorithmic solution that makes worked examples more effective than problem-solving. This can be tested by using worked examples that only consist of a problem statement and solution, or poorly elaborated solution steps. In such cases, worked examples alone might not be as effective as in combination with self-explanations. In addition, collecting think aloud data in a study that is experimentally set up like the one we have reported will further highlight evidence of cognitive processes learners engage in and can be used to

validate the theoretical interpretation of our findings. Further, our analyses did not shed light on whether learners who self-explained one problem solution perfectly and did not solve any of the other two problems differ in their learning outcome from learners who self-explained, say, one part of each of the three problems perfectly, but no problem fully. Whereas the latter case might reflect a perfect self-explanation of one solution step across each of the problems, but a failure to learn about the other two relevant solution steps, the former case might have processed the full solution procedure which, although only self-explained once, would already lead to better understanding of the strategy and consequently, application of the procedure to new problems. Such an analysis could shed further light on whether worked examples have this direct effect of conveying a problem solution procedure or strategy that we described above.

Further, it would be interesting to systematically investigate whether self-explaining helps learners with lower prior knowledge to benefit more from problem-solving as this learning activity might potentially alleviate the disadvantage problem-solving posits for these learners. Likewise, it seems worthwhile to investigate whether there is a qualitative difference in self-explanations if learners are asked to explain an example versus their own problem solution (i.e., their application of concepts to specific problem information). Future research could determine whether such differences exist or matter. In the same vein, it might be interesting to investigate if the self-explanation instruction determines what kind of self-explanations are elicited. For example, whether in procedural and nonprocedural domains the prompt to "justify" versus the prompt to "self-explain" elicits the same kind of explanations.

Conclusion

We reported the analyses of Schwaighofer et al.'s (2016) self-explanation data and used conditional process analysis (Hayes, 2018) to jointly investigate moderating and mediating factors in a worked examples—problem-solving paradigm. We discussed different approaches of conceptualizing self-explanation quality and identified that numerous features including the types of problems, the domain, or the prompts used determine the kind of self-explanations that will be elicited or the quality aspects that can be observed. As such, we propose to incorporate reflections on what specific features of a study mean for the generalizability of the results in future self-explanation research. Further, we propose to measure justification of claims to assess self-explanation quality for ill-defined problems or when learners basically apply conceptual information to specific problems.

Based on our analyses, self-explanation quality does not seem to mediate the effect of worked examples on learning gains in a worked examples—problem-solving paradigm. We may not have captured what explains the worked example effect. Different kinds of self-explanations or different aspects of quality may mediate effects but were not detected with our coding approach or were simply not reflected in the artifacts our learners created. Yet, the results of the present study may serve as a reason to pause and reflect on the available evidence for this presumed mediation effect. First, studies that found mediation effects often tested the effect of prompting self-explanations and students often explained different content in the experimental and control conditions. Second, studies that compared different types of examples or worked examples and problem-solving are just beginning to apply state-of-the-art mediation analysis methods. Third, these studies usually have low statistical power (including our study) for detecting moderation or mediation effects.

Based on our results we thus tentatively suggest that worked examples have a direct effect: In the case of ill-defined (statistics) problems, worked examples convey a solution strategy which is more important for successful future problem-solving than justifying claims in solutions to practice problems. Alternatively, examples prepare learners to use analogical reasoning when solving new problems. Cumulative evidence from highly powered studies is needed to corroborate these explanations. Our study aligns with other evidence showing that self-explaining worked examples supports learners in actively processing the examples (Rittle-Johnson, 2006). It also suggests that self-explained problem-solving is more effective than pure problem-solving. Thus, the instructional value of self-explanation remains even though this study found no evidence for a mediating role of self-explanation quality in the worked examples—problem-solving paradigm. Finally, investigating moderating and mediating variables jointly advances our understanding of the worked example effect and by extension instructional effectiveness.

Acknowledgements We received funding from the Elite Network of Bavaria Grant K-GS-2012-209.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Adams, D. M., McLaren, B. M., Durkin, K., Mayer, R. E., Rittle-Johnson, B., Isotani, S., & van Velsen, M. (2014). Using erroneous examples to improve mathematics learning with a web-based tutoring system. *Computers in Human Behavior, 36*, 401–411. https://doi.org/10.1016/j.chb.2014.03.053

Aguinis, H. (1995). Statistical power with moderated multiple regression in management research. *Journal of Management, 21*(6), 1141–1158. https://doi.org/10.1177/014920639502100607

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90*(1), 94–107. https://doi.org/10.1037/0021-9010.90.1.94

Allwood, C. M. (1984). Error detection processes in statistical problem solving. *Cognitive Science, 8*(4), 413–437. https://doi.org/10.1207/s15516709cog0804_5

Allwood, C. M., & Montgomery, H. (1982). Detection of errors in statistical problem solving. *Scandinavian Journal of Psychology, 23*(1), 131–139. https://doi.org/10.1111/j.1467-9450.1982.tb00423.x

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*. 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. https://doi.org/10.3102/00346543070002181

Berthold, K., Eysink, T. H., & Renkl, A. (2009). Assisting self-explanation prompts are more effective than open prompts when learning with multiple representations. *Instructional Science, 37*(4), 345–363. https://doi.org/10.1007/s11251-008-9051-z

Berthold, K., & Renkl, A. (2009). Instructional aids to support a conceptual understanding of multiple representations. *Journal of Educational Psychology, 101*(1), 70–87. https://doi.org/10.1037/a0013247

Bichler, S., Schwaighofer, M., Stadler, M., Bühner, M., Greiff, S., & Fischer, F. (2020). How working memory capacity and shifting matter for learning with worked examples: A replication study. *Journal of Educational Psychology, 112*(7), 1320–1337. https://doi.org/10.1037/edu0000433

Chi, M. T. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in Instructional Psychology* (pp. 161–238). Lawrence Erlbaum Associates.

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*(2), 145–182. https://doi.org/10.1207/s15516709cog1302_1

Chi, M. T., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439–477. https://doi.org/10.1207/s15516709cog1803_3

Chi, M. T., Slotta, J. D., & De Leeuw, N. (1994). From things to processes: A theory of conceptual change for learning science concepts. *Learning and Instruction, 4*(1), 27–43. https://doi.org/10.1016/0959-4752(94)90017-5

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

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. https://doi.org/10.1037/dev0000075

Friedman, N. P., Miyake, A., Robinson, J. L., & Hewitt, J. K. (2011). Developmental trajectories in toddlers' self-restraint predict individual differences in executive functions 14 years later: A behavioral genetic analysis. *Developmental Psychology, 47*(5), 1410–1430. https://doi.org/10.1037/a0023750

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

Hefter, M. H. (2021). Web-based training and the roles of self-explaining, mental effort, and smartphone usage. *Technology, Knowledge and Learning*. https://doi.org/10.1007/s10758-021-09563-w

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. https://doi.org/10.1007/s11251-014-9320-y

Hefter, M. H., Renkl, A., Riess, W., Schmid, S., Fries, S., & Berthold, K. (2015). Effects of a training intervention to foster precursors of evaluativist epistemological understanding and intellectual values. *Learning and Instruction, 39*, 11–22. https://doi.org/10.1016/j.learninstruc.2015.05.002

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

Lakens, D., & Evers, E. R. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science, 9*(3), 278–292. https://doi.org/10.1177/1745691614528520

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. https://doi.org/10.1177/0963721411429458

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. https://doi.org/10.1006/cogp.1999.0734

Ohlsson, S. (1996). Learning from performance errors. *Psychological Review, 103*(2), 241–262. https://doi.org/10.1037/0033-295X.103.2.241

Paas, F., & Sweller, J. (2014). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning*. 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. https://doi.org/10.1037/0022-0663.84.4.429

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. https://doi.org/10.1037/0022-0663.86.1.122

Preacher, K. J., & Sterba, S. K. (2019). Aptitude-by-treatment interactions in research on educational interventions. *Exceptional Children, 85*(2), 248–264. https://doi.org/10.1177/0014402918802803

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment, 28*(3), 164–171. https://doi.org/10.1027/1015-5759/a000123

Reimann, P., & Schult, T. J. (1996). Turning examples into cases: Acquiring knowledge structures for analogical problem solving. *Educational Psychologist, 31*(2), 123–132.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science, 21*(1), 1–29. https://doi.org/10.1207/s15516709cog2101_1

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cognitive Science, 38*(1), 1–37. https://doi.org/10.1111/cogs.12086

Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: The effects of example variability and elicited self-explanations. *Contemporary Educational Psychology, 23*(1), 90–108. https://doi.org/10.1006/ceps.1997.0959

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development, 77*(1), 1–15. https://doi.org/10.1111/j.1467-8624.2006.00852.x

Rittle-Johnson, B., & Loehr, A. M. (2017). Eliciting explanations: Constraints on when self-explanation aids learning. *Psychonomic Bulletin & Review, 24*(5), 1501–1510. https://doi.org/10.3758/s13423-016-1079-5

Rittle-Johnson, B., Loehr, A. M., & Durkin, K. (2017). Promoting self-explanation to improve mathematics learning: A meta-analysis and instructional design principles. *ZDM Mathematics Education, 49*(4), 1–13. https://doi.org/10.1007/s11858-017-0834-z

Roelle, J., & Renkl, A. (2020). Does an option to review instructional explanations enhance example-based learning? It depends on learners' academic self-concept. *Journal of Educational Psychology, 112*(1), 131–147. https://doi.org/10.1037/edu0000365

Roy, M., & Chi, M. T. (2005). The self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 271–286). Cambridge University Press.

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science, 8*(4), 379–386. https://doi.org/10.1177/1948550617715068

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. https://doi.org/10.1007/s10648-011-9168-5

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. https://doi.org/10.1037/edu0000115

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

Schworm, S., & Renkl, A. (2007). Learning argumentation skills through the use of prompts for self-explaining examples. *Journal of Educational Psychology, 99*(2), 285–296. https://doi.org/10.1037/0022-0663.99.2.285

Stark, R., Gruber, H., Renkl, A., & Mandl, H. (2000). Instruktionale Effekte einer kombinierten Lernmethode. Zahlt sich die Kombination von Lösungsbeispielen und Problemlöseaufgaben aus? *Zeitschrift Für Pädagogische Psychologie, 14*(4), 206–218. https://doi.org/10.1024//1010-0652.14.4.206

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7

Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*, 261–292. https://doi.org/10.1007/s10648-019-09465-5

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

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

VanLehn, K., & Jones, R. M. (1993). What mediates the self-explanation effect? Knowledge gaps, schemas or analogies? In M. Polsen (Ed.), *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 1034–1039). Erlbaum

Authors and Affiliations

Sarah Bichler[1,3] · Matthias Stadler[1] · Markus Bühner[1] · Samuel Greiff[2] · Frank Fischer[1]

✉ Sarah Bichler
sbichler@berkeley.edu

[1] Department of Psychology, Ludwig-Maximilans-Universität München, Munich, Germany

[2] Institute of Cognitive Science and Assessment, Université du Luxembourg, Esch-sur-Alzette, Luxembourg

[3] Present Address: School of Education, University of California, Berkeley, 2121 Berkeley Way, Berkeley, CA 94720-1670, USA