Metacognition and Learning (2022) 17:565–588
https://doi.org/10.1007/s11409-022-09293-z

Expert example standards but not idea unit standards help learners accurately evaluate the quality of self-generated examples

Linda Froese[1] · Julian Roelle[1]

Received: 14 June 2021 / Accepted: 9 February 2022 / Published online: 24 March 2022

Abstract

Generating own examples for previously encountered new concepts is a common and highly effective learning activity, at least when the examples are of high quality. Unfortunately, however, students are not able to accurately evaluate the quality of their own examples and instructional support measures such as idea unit standards that have been found to enhance the accuracy of self-evaluations in other learning activities, have turned out to be ineffective in example generation. Hence, at least when learners generate examples in self-regulated learning settings in which they scarcely receive instructor feedback, they cannot take beneficial regulation decisions concerning when to continue and when to stop investing effort in example generation. The present study aimed at investigating the benefits of a relatively parsimonious means to enhance judgment accuracy in example generation tasks, i.e. the provision of expert examples as external standards. For this purpose, in a 2×2 factorial experiment we varied whether $N = 131$ university students were supported by expert example standards (with vs. without) and idea unit standards (with vs. without) in evaluating the quality of self-generated examples that illustrated new declarative concepts. We found that the provision of expert example standards reduced bias and enhanced absolute judgment accuracy, whereas idea unit standards had no beneficial effects. We conclude that expert example standards are a promising means to enhance judgment accuracy in evaluating the quality of self-generated examples.

Keywords example generation · metacognition · monitoring · judgment accuracy · overconfidence · standards

✉ Linda Froese
linda.froese@ruhr-uni-bochum.de

[1] Faculty of Philosophy and Educational Research, Ruhr University Bochum, Universitätsstraße 150, 44801 Bochum, Germany

Generating own examples for previously encountered content is a common generative learning activity. For instance, in learning by teaching (e.g., Fiorella & Mayer, 2013; for an overview, see Lachner et al., 2021), in self-regulated learning via journal writing (e.g., Moning & Roelle, 2021; for an overview see Nückles et al., 2020), or in self-explaining (e.g., Wylie & Chi, 2014, for a recent meta-analysis see Bisra et al., 2018), learners, amongst other elaborative activities, usually generate examples that illustrate new principles and concepts that are to be learned. Furthermore, example generation also serves as a beneficial stand-alone activity. A series of experiments by Rawson and Dunlosky (2016) shows that generating own examples for to-be-learned declarative concepts is more effective than engaging in shallow activities such as restudy.

To be beneficial in terms of learning outcomes, however, the generated examples need to be of high quality. Because the time and resources required make it unlikely that learners will receive sufficient instructor feedback on the quality of their examples in most tasks that involve example generation, it is therefore critical that learners are able to accurately evaluate the quality of self-generated examples themselves. More specifically, this ability would be crucial for learners to take beneficial regulation decisions, that is to decide when to stop and when to continue investing effort in generating examples regarding the content that is to be learned. Unfortunately, however, a study by Zamary et al. (2016) clearly indicates that learners' ability in doing so is poor. Specifically, the authors found evidence for substantial overconfidence and even the provision of idea unit standards, which has shown to foster judgment accuracy in other learning tasks (e.g., Lipko et al., 2009), did not enhance the accuracy of learners' evaluations.

Against this background, the present study was designed to discover means to reduce the outlined student inaccuracy in evaluating self-generated examples. Specifically, in an attempt to replicate and extend the findings by Zamary et al. (2016), we factorially varied whether university students were supported by idea unit standards (with vs. without) and newly developed expert example standards (with vs. without) in evaluating their examples. Three measures of judgment accuracy (i.e., bias, absolute accuracy, and relative accuracy) were used as the main dependent variables.

Example Generation and Evaluation: Why it Matters

In supporting learners to acquire both basic declarative concepts, which are part of textbooks in almost any content domain, and advanced complex principles, a widespread and effective means is engaging learners in generative learning activities such as elaboration (e.g., Fiorella & Mayer, 2016; see also Brod, 2021). A particularly widespread elaborative activity, which has proven its effectiveness both as a stand-alone activity and as part of learning tasks that trigger various generative activities such as learning by teaching (e.g., Hoogerheide et al., 2019; Lachner et al., 2021), journal writing (e.g., Berthold et al., 2007; Nückles et al., 2020), or self-explaining (e.g., Bisra et al., 2018; Wylie & Chi, 2014), is to generate own examples. In generating own examples, learners generate meaningful new information based on the given information and prior knowledge. This activity is theorized to enrich learners' mental representations of the to-be-learned content and integrate the mental representations with existing prior knowledge, which finally improves comprehension (Fiorella & Mayer, 2016, see also Weinstein & Mayer, 1986).

Importantly, the benefits of example generation depend on example quality. More specifically, similar to most generative learning activities, the benefits of example generation increase with increasing quality (e.g., Glogger et al., 2012; Rawson & Dunlosky, 2016). Hence, in order to optimize the benefits of example generation, it is crucial that example quality is accurately monitored. Due to the involved time and resources, however, both in settings in which learners engage in example generation in a self-regulated manner (e.g., Gurung et al., 2010; see also Roelle et al., 2017b; Nückles et al., 2020) and in settings in which instructors engage learners in generating examples as part of in-class or homework assignments, it is unlikely that learners receive sufficient instructor feedback on their examples (cf. Roelle et al., 2011). Hence, to be able to take beneficial regulation decisions in generating examples (e.g., concerning when to stop and when to continue investing effort in generating quality examples), it is crucial that learners can evaluate the quality of their examples on their own.

Unfortunately, however, learners' ability to accurately evaluate self-generated examples is low. In two experiments with university students, Zamary et al. (2016) found consistent evidence for substantial inaccuracy (mostly: overconfidence). Although inaccurate judgments of learning are the rule rather than the exception in various task assignments (see e.g., De Bruin et al., 2020; Prinz et al., 2020), what makes this finding exceptional is that an established means to enhance judgment accuracy did not help. Specifically, providing learners with idea unit and full standards scarcely affected judgment accuracy in evaluating self-generated examples.

Example Evaluation: Why Idea Unit Standards Might not Help, but Expert Example Standards Might

Motivated by the *absence of standards hypothesis*, which states that when learners do not have access to the objectively correct response to a task (i.e., an external standard), they have difficulties in accurately evaluating their own responses, standards have been introduced as a means to enhance judgment accuracy (see Rawson & Dunlosky, 2007). Basically, standards are external representations of a correct answer to a task assignment. The provision of standards can substantially enhance judgment accuracy. For example, Rawson and Dunlosky (2007) showed that college students were better able to assess their own performance in retrieval practice tasks when they were provided with the correct answers to the tasks as standards. These results could be replicated for middle school students by Lipko et al. (2009) and extended to problem-solving tasks by Baars et al. (2014). Lipko and colleagues also further developed the format of the standards. Specifically, the authors developed so-called *idea unit standards* by dividing the correct answer to the respective task into its constituent idea units and could show that this format, potentially because it highlights the crucial components of the target response more clearly, has added value to providing learners with *full standards* that are not divided into idea units. A further benefit of idea unit standards could be that they are mindful of learners' working memory capacity. Idea unit standards allow to evaluate learner responses concerning one idea unit at a time (see Dunlosky et al., 2011). As learners' cognitive resources are limited (e.g., Sweller et al. 2019), this potential offloading function of idea unit standards in forming self-evaluations might contribute to the benefits of idea unit standards as well.

In view of the robust evidence for the benefits of idea unit standards, Zamary et al. (2016) investigated whether idea unit standards (as well as *full standards* that were not divided into idea units) would also exert beneficial effects on judgment accuracy in example generation tasks. For this purpose, they had learners study definitions of new declarative concepts and then generate an own example for each concept. Afterwards, learners were to evaluate their examples by predicting whether the examples would receive either no credit, partial credit or full credit when they were graded. In forming their evaluations, some of the learners were supported by idea unit standards or by full standards. The idea unit standards were generated by breaking down the concept definitions into their constituent parts and the learners were instructed to judge whether the idea units were adequately illustrated by their self-generated examples before predicting whether their examples would receive no, partial, or full credit; the full standards were basically the concept definitions. The authors found that neither idea unit standards nor full standards enhanced the accuracy of learners' self-evaluations and that, in all conditions, learners were largely overconfident in evaluating their examples.

One explanation for the lack of beneficial effects of idea unit and full standards in example generation tasks is that, other than in the studies on retrieval practice tasks and problem-solving tasks (e.g., Baars et al., 2014; Lipko et al., 2009; Rawson & Dunlosky, 2007), the standards in Zamary et al.'s study did not provide learners with access to objectively correct responses to the tasks. Obviously, the constituent idea units of a concept definition or the concept definitions themselves are not a correct example that illustrates the concept definition. Hence, other than in the outlined previous studies, the learners could not directly compare their products (i.e., the generated examples) with the provided standards but, just like in generating the examples in the first place, had to generate an internal standard for what would constitute a correct illustration of each idea unit/each definition and then evaluate whether their examples would accurately illustrate the respective idea units/definitions. The internal standards, however, can be misaligned with objective standards. Specifically, in view of the finding that even when external standards are well aligned with learners' products and hence direct comparisons are possible, learners often miss substantial amounts of inconsistencies between their products and the (idea unit) standards (e.g., Lipko et al., 2009; Rawson & Dunlosky, 2007), it is reasonable to assume that effectively using idea unit or full standards in evaluating own examples was beyond learners' competence. This notion is supported by Zamary et al.'s finding that, on average, the examples received objective ratings of only ca. 40% - when learners are not able to generate quality examples on their own, they likely are also not able to form accurate internal standards based on provided idea unit or full standards and hence to accurately evaluate their examples. Consequently, the standards did not increase the accuracy of learners' self-evaluations.

As an alternative standard in example generation tasks, which would actually provide learners with objectively correct responses to the tasks, expert examples could be used. Research on learning from (worked) examples clearly indicates that learners are usually well able to relate concrete examples to previously studied abstract principles and concepts (e.g., Berthold & Renkl, 2009; Roelle et al., 2017a; Zamary & Rawson, 2018; for an overview, see Renkl, 2014). Hence, it is reasonable to assume that when provided with expert example standards, learners would be able to see how the examples illustrate the respective concepts. This, in turn, should enable them to use the expert examples as standards for an objectively sufficient illustration of the respective concepts and hence to engage in meaningful (relatively direct) comparisons of their own examples with the expert examples that finally result in increased judgment accuracy. Hence, it could be assumed that, other than idea unit or full standards, expert example standards exert beneficial effects on judgment accuracy in example generation tasks. In support of this notion, findings by Waldeyer and Roelle (2021) suggest that expert responses can serve as a beneficial standard in determining the quality of self-generated keywords for previously read texts. However, generating keywords for previously encountered material arguably resembles retrieval practice tasks more than example generation tasks; furthermore, Waldeyer and Roelle did not assess established judgment accuracy measures. Therefore, while these findings suggest the potential of expert example standards in enhancing judgment accuracy, they by no means already indicate that the desired effects on judgment accuracy in example generation tasks are likely to occur.

A cautious prediction concerning the potential benefits of expert example standards is furthermore warranted because, in comparison to idea unit or full standards in retrieval practice or problem-solving tasks, expert example standards in example generation tasks arguably require more interpretation. In most cases, the surface features of the expert examples (e.g., the cover stories) would be different from the surface features of learners' own examples and, accordingly, learners would have to transfer, based on the way the concepts are illustrated in the cover stories of the expert examples, what a similarly good illustration would have to look like in their cover stories. This transfer could prove difficult for learners, which could correspondingly increase cognitive load in self-evaluating their own examples. As learners who perform a task such as generating own examples for new concepts for the first time usually experience substantial cognitive load while performing the task, which can leave little to no working memory capacity for evaluating their task performance (e.g., Panadero et al., 2016; see also Kostons et al., 2009, 2012), the outlined difficulty could potentially overtax learners. If processing the expert example standards and the subsequent forming of self-evaluations would be too difficult and hence overload learners, no beneficial effects on judgment accuracy could be expected. Furthermore, in this case a reduction of the mental effort that learners invest into processing the expert example standards and forming self-evaluations could be expected, because learners might not consider processing the expert examples as a beneficial investment of their cognitive resources (see Schnotz, 2010).

In addition to investigating the effects of expert example standards on judgment accuracy, it could thus also be fruitful to assess measures of subjective cognitive load such as perceived task difficulty and mental effort in processing the expert example standards and the subsequent forming of self-evaluations. The construct of perceived task difficulty indicates the level of difficulty that learners experience during performing a task, whereas mental effort describes the amount of controlled resources learners allocate to performing a task (see van Gog & Paas, 2008). Although these measures are merely subjective appraisals that could be biased by several factors (for an elaborate discussion, see Scheiter et al., 2020), they could prove informative when expert example standards would not entail beneficial effects on judgment accuracy in particular. If a lack of effect of expert example standards on judgment accuracy would go along with high subjective difficulty and low invested mental effort in processing the standards and forming self-evaluations, it would suggest that designing means to reduce cognitive load in processing expert example standards would be a sensible next step in exploring the benefits of expert example standards.

The Present Study

In view of these empirical and theoretical considerations, the main goal of the present study was to investigate the effects of expert example standards in enhancing judgment accuracy in evaluating self-generated examples. To strengthen the notion that idea unit standards are not effective in this manner, in an attempt to replicate the findings of Zamary et al. (2016), we investigated the effects of idea unit standards as well. In addition to investigating how accurate learners can evaluate the quality of self-generated examples (RQ 1), our main research questions were whether the provision of idea unit standards (RQ 2a) and expert example standards (RQ 2b) would enhance students' judgment accuracy in evaluating the quality of self-generated examples. In view of the notion that expert example standards could potentially overload learners, which would prevent effects of judgment accuracy to occur, we also investigated whether expert example standards (and idea unit standards as well) would affect subjective task difficulty and mental effort invested in processing the standards and evaluating the quality of self-generated examples (RQ 3a and 3b, respectively). For explorative purposes, we also analyzed whether idea unit and expert example standards would interact in terms of judgment accuracy as well as subjective task difficulty and mental effort.

Method

Sample and Design

As there were hardly any previous studies that analyzed the effects of expert example standards in evaluating the quality of self-generated examples, which was the focal innovative support measure in the present study, we used a medium effect size ($\eta_p^2 = .06$; further parameters: $\alpha = .05$, $\beta = .20$) as the basis for our a priori power analysis. The power analysis was conducted with G*Power 3.1.9.3 (Faul et al., 2007). A 2×2 ANOVA, which corresponded with the study's design (see below), was used as the statistical test for which the required sample size was determined. The analysis yielded a required sample size of $N = 128$. Against this background, we recruited $N = 131$ university students (101 female, 30 male; $M_{Age} = 24.08$ years, $SD_{Age} = 3.93$ years) who attended different universities in Germany. They received € 10 for their participation. Written informed consent was collected from all participants.

As the study was designed in part as a replication of the study by Zamary et al. (2016), the procedure was identical to Zamary et al.'s study except for minor adjustments detailed below. Like in the study by Zamary and colleagues, after an initial study phase, in which all learners first read a short expository text that covered eight declarative concepts about social attribution and then were presented with the eight concepts again one at a time together with the corresponding definition, all participants were prompted to generate own examples for each of the concepts. They then were asked to evaluate the quality of their examples by assigning no, partial or full credit. The experimental manipulation was carried out in the evaluation phase. We factorially varied whether the learners were supported by *idea unit standards* (with vs. without) and (b) *expert example standards* (with vs. without) in evaluating their examples, which resulted in a 2×2 factorial between-subject design. The participants were randomly assigned to the four experimental conditions.

Materials

Expository Text and Concept Definitions

The expository text that all learners were instructed to read carefully in the initial study phase covered eight declarative concepts about social attribution (for an overview of the

Table 1 Overview of the Declarative Concepts (Translated From German)*.

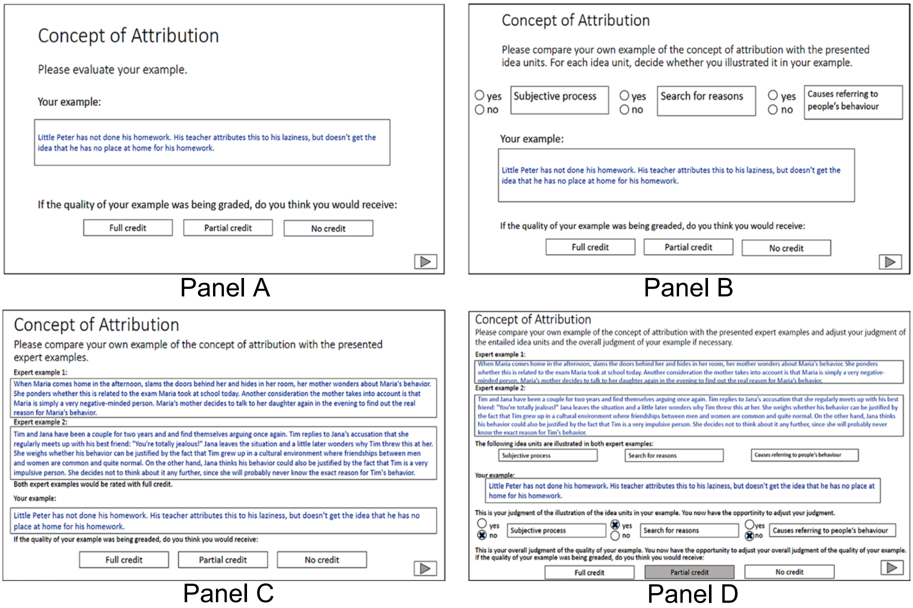
Term	Full Definition
Attribution	The process through which we seek to explain certain behaviors or events.
Social Norms	Explicit or implicit conventions that dictate appropriate behavior in social situations.
Consensus	The extent to which behavior by one person is shown by others as well. The consensus is high, when a lot of people react in the same way, and low, when only a few people react in the same way.
Consistency	The extent to which a person exhibits similar behavior in response to a given stimulus or situation. The consistency is high, when the behavior is similar over a period of time, and low, when the behavior remains the same only for a short time.
Distinctiveness	The extent to which a person reacts in the same manner to different stimuli or situations. The distinctiveness is high, when a person behaves the same in only a few situations, and low, when the person shows the same behavior in many similar situations.
Correspondence Bias	The tendency to attribute other people's behavior to internal causes to a greater extent than is actually justified while underestimating the effect of the situation.
Self-Serving Bias	The tendency to attribute positive outcomes to our own traits or characteristics (internal causes) but negative outcomes to factor beyond our control (external causes).
Just-World Hypothesis	The strong desire or need people have to believe that the world is an orderly, predictable, and just place, where people get what they deserve.

*The wording was adapted from the material used by Zamary et al. (2016).

concepts, see Table 1). Specifically, in order to be able to meaningfully relate the present study's results to the study of Zamary et al. (2016), we used a translated and slightly adapted version of the expository text that was used by Zamary and colleagues. The text was comprised of 476 words and did not include any examples. Yet the text included not only the definitions, but also introduced the learners to the overall topic of social cognition and provided explanations on how the concepts are associated with general human behavior (e.g., the concept of attribution was contextualized in the notion that people frequently question other peoples' behavior in everyday life). After the students had finished reading the text, similar to the procedure of Zamary et al. (2016), the eight concepts were presented one at a time with the corresponding definitions (in the same sequence as they were explained in the expository text). The learners were asked to carefully read the definitions. The expository text and concept definitions can be viewed under https://doi.org/10.17605/OSF.IO/2BSP6.

Support Measures in Evaluating the Examples

After the learners had carefully studied the expository text and the concept definitions, they were prompted to generate an example for each concept (one at a time). Like in Zamary et al. (2016), the respective concept definition was visible during the generation of the example (*open-book* generative task, see Blunt & Karpicke, 2014; Roelle & Berthold,



Concept of Attribution

Please evaluate your example.

Your example:

Little Peter has not done his homework. His teacher attributes this to his laziness, but doesn't get the idea that he has no place at home for his homework.

If the quality of your example was being graded, do you think you would receive:

Full credit | Partial credit | No credit

Panel A

Concept of Attribution

Please compare your own example of the concept of attribution with the presented idea units. For each idea unit, decide whether you illustrated it in your example.

○ yes ○ no Subjective process

○ yes ○ no Search for reasons

○ yes ○ no Causes referring to people's behaviour

Your example:

Little Peter has not done his homework. His teacher attributes this to his laziness, but doesn't get the idea that he has no place at home for his homework.

If the quality of your example was being greaded, do you think you would receive:

Full credit | Partial credit | No credit

Panel B

Concept of Attribution

Please compare your own example of the concept of attribution with the presented expert examples.

Expert example 1:

When Maria comes home in the afternoon, slams the doors behind her and hides in her room, her mother wonders about Maria's behavior. She ponders whether this is related to the exam Maria took at school today. Another consideration the mother takes into account is that Maria is simply a very negative-minded person. Maria's mother decides to talk to her daughter again in the evening to find out the real reason for Maria's behavior.

Expert example 2:

Tim and Jana have been a couple for two years and and find themselves arguing once again. Tim replies to Jana's accusation that she regularly meets up with his best friend: "You're totally jealous!" Jana leaves the situation and a little later wonders why Tim threw this at her. She weighs whether his behavior can be justified by the fact that Tim grew up in a cultural environment where friendships between men and women are common and quite normal. On the other hand, Jana thinks his behavior could also be justified by the fact that Tim is a very impulsive person. She decides not to think about it any further, since she will probably never know the exact reason for Tim's behavior.

Both expert examples would be rated with full credit.

Your example:

Little Peter has not done his homework. His teacher attributes this to his laziness, but doesn't get the idea that he has no place at home for his homework.

If the quality of your example was being greaded, do you think you would receive:

Full credit | Partial credit | No credit

Panel C

Concept of Attribution

Please compare your own example of the concept of attribution with the presented expert examples and adjust your judgment of the entailed idea units and the overall judgment of your example if necessary.

Expert example 1:

When Maria comes home in the afternoon, slams the doors behind her and hides in her room, her mother wonders about Maria's behavior. She ponders whether this is related to the exam Maria took at school today. Another consideration the mother takes into account is that Maria is simply a very negative-minded person. Maria's mother decides to talk to her daughter again in the evening to find out the real reason for Maria's behavior.

Expert example 2:

Tim and Jana have been a couple for two years and find themselves arguing once again. Tim replies to Jana's accusation that she regularly meets up with his best friend: "You're totally jealous!" Jana leaves the situation and a little later wonders why Tim threw this at her. She weighs whether his behavior can be justified by the fact that Tim grew up in a cultural environment where friendships between men and women are common and quite normal. On the other hand, Jana thinks his behavior could also be justified by the fact that Tim is a very impulsive person. She decides not to think about it any further, since she will probably never know the exact reason for Tim's behavior.

The following idea units are illustrated in both expert examples:

Subjective process | Search for reasons | Causes referring to people's behaviour

Your example:

Little Peter has not done his homework. His teacher attributes this to his laziness, but doesn't get the idea that he has no place at home for his homework.

This is your judgment of the illustration of the idea units in your example. You now have the opportinity to adjust your judgment.

○ yes ☒ no Subjective process

☒ yes ○ no Search for reasons

○ yes ☒ no Causes referring to people's behaviour

This is your overall judgment of the quality of your example. You now have the opportunity to adjust your overall judgment of the quality of your example.
If the quality of your example was being greaded, do you think you would receive:

Full credit | [✓ Partial credit] | No credit

Panel D

Fig. 1 Example-quality evaluation screens for the concept of attribution for each group. (Panel A) No Feedback, (Panel B) Idea Unit Standards, (Panel C) Expert Example Standards, (Panel D) Combination of Idea Unit Standards and Expert Example Standards (this group was provided with two screens: screen 1 that is similar to Panel B, and screen 2, shown here, entailing the ratings made on screen 1).

2017; Waldeyer et al., 2020).[1] Specifically, they received the prompt "Please generate an example that illustrates the concept of […]." Immediately after each generation trial, the learners were asked to evaluate the quality of their example. For this purpose, we used a translated version of the instruction that was used in Zamary et al. (2016): "If the quality of your example was being graded, do you think you would receive: no credit, partial credit, or full credit?"

When evaluating the examples, the condition without idea unit standards and without expert example standards was given only the concept term (see Panel A of Fig. 1). The idea unit standards condition was given the concept term alongside with the definition of the concept that was broken down in idea units. We used the idea units of Zamary et al. (2016) as a basis, but implemented some slight adaptations. Specifically, in an attempt to simplify evaluating whether the respective idea units were or were not covered by an example, we partly aggregated idea units that were unclear when standing alone (e.g., for the concept of correspondence bias, the idea units *the tendency to attribute* and *other people's behavior* were converted into one idea unit *attribution concerning other people's behavior*). These adaptations resulted in two concepts with two idea units, four concepts with three idea units, and two concepts with four idea units (in Zamary et al., it was four concepts with three idea units, two concepts with four idea units, one concept with five idea units, and one concept with six idea units; the idea units can be viewed under https://doi.org/10.

[1] Zamary and colleagues also implemented a condition in which the concept definitions were not visible during example generation. For the present study, however, this closed-book condition is not relevant.

17605/OSF.IO/2BSP6). Like in Zamary et al., the students were asked to check for each idea unit whether it was or was not illustrated in their example before making an overall judgment of the quality of their example by assigning no, partial or full credit (see Panel B of Fig. 1).

The expert example standards condition received two expert examples per concept (i.e., 16 expert examples in total), which were generated such that they would receive full credit. More specifically, to ensure sufficient quality the expert examples were piloted with four research assistants who were highly knowledgeable concerning the eight declarative concepts and who rated whether all idea units of the respective concepts were correctly illustrated in the examples. Only examples that were considered as illustrating all idea units correctly by all experts were used in the present study. During the example evaluation, the expert example standards were presented above the self-generated example, with the information that the two expert examples would receive full credit. The learners were asked to compare the expert examples with their own examples and then rate whether their examples would receive no, partial, or full credit (see Panel C of Fig. 1).

The learners who received both support measures first received the idea unit standards (identical to the idea unit standards condition). Then, on the next page, they were shown the expert example standards on top of the page as well as their judgments regarding the covered idea units and the overall judgment on the bottom of the page while being provided with the opportunity to revise these two assessments they made previously (see Panel D of Fig. 1).

Instruments and Measures

Assessment of Academic Self-Concept

As it is one of the strongest motivational predictors for learner performance (e.g., Lotz et al. 2018; Steinmayr & Spinath, 2009), we measured learners' academic self-concept as a control variable. The academic self-concept is an ability-related self-appraisal that is defined as a learner's self-perception of his or her academic abilities and competencies (e.g., Byrne & Shavelson, 1986). The questions were adapted from the absolute self-concept scale of the SESSKO (Schöne et al., 2002; e.g., "*For my studies, I am*… 1: *not talented* – 5: *very talented*"). The items were scored on 5-point Likert scales and aggregated for the subsequent analyses (Cronbach's α = .81).

Pretest

As it can be a strong cognitive predictor for learner performance (e.g., Simonsmeier et al., 2021), we measured learners' prior knowledge as a further control variable. Specifically, we assessed learners' prior knowledge regarding the eight concepts that were to be learned in the present study with a pretest that asked them to write down the definitions of the declarative concepts one at a time ("Please define the concept of […]"). The answers to the eight questions were scored by two independent raters who were blind to the experimental conditions. In particular, they determined for each idea unit whether it was or was not covered in the answers. On this basis, no, partial or full credit was assigned for each question. Interrater reliability between the two raters, measured by the intraclass correlation coefficient with measures of absolute agreement, was very good for all questions (all

ICCs > .85). For the later analyses, the scores of all eight pretest items were converted into percentages and then averaged (i.e., 0–100%; Cronbach's α = .50).

Assessment of Example Quality

To assess the quality of the students' self-generated examples, based on a scoring protocol two independent raters who were blind to the experimental conditions evaluated whether the respective idea units of the concepts were correctly illustrated in the learners' examples. For instance, one participant generated the following example to illustrate the concept of self-serving bias: "Fabian participated in a sports bet and lost a lot of money. He thinks it's because the soccer team he bet on played so badly." The concept of self-serving bias was split into the two idea units *tendency to attribute positive outcomes to our own characteristics (internal causes)* and *tendency to attribute negative outcomes to factors beyond our control (external causes)*. The respective example was rated with partial credit in total, because it illustrated the idea unit *tendency to attribute negative outcomes to factors beyond our control (external causes)*, whereas the idea unit *tendency to attribute positive outcomes to our own characteristics (internal causes)* was not illustrated. In contrast, the following example that illustrates the concept of self-serving bias was assigned with full credit: "In the math test, Sarah receives a very good grade, which she justifies by saying that she is simply very talented in math. In the German test, on the other hand, she gets a poor grade, which she believes can only be due to the teacher's unfair evaluation." The assignment of full credit is due to the fact that this example correctly illustrated both of the outlined idea units. Based on these evaluations, the raters determined whether the examples were assigned with no, partial or full credit. Interrater reliability was very good for the examples regarding all of the eight declarative concepts (all ICCs > .85). For the later analyses, the learners' scores were averaged across all eight concepts (Cronbach's α = .65) and transformed to percentage scores (i.e., 0–100%).

Judgment Accuracy

To determine the accuracy of learners' evaluations of their examples, we used the learners' ratings and the expert ratings. Specifically, based on these ratings we formed three different measures that describe judgment accuracy according to Schraw (2009): *bias*, *absolute accuracy* and *relative accuracy*. First, we computed bias scores by subtracting the experts' from the students' ratings (like the expert ratings, these were converted into percentages beforehand). Hence, positive and negative results were possible, indicating overconfidence for positive values, underconfidence for negative values and accurate judgments for the score zero (i.e., values between -100% and 100% were possible).

One limitation of the bias measure, however, is that when there are underconfident and overconfident students within one or multiple condition(s) or ratings within one student, over- and underconfident ratings may cancel each other out and result in less robust judgment accuracy estimates. Therefore, we also calculated the absolute deviation scores, referred to as *absolute accuracy*, which is defined as the degree of correspondence between a learner's judgment and her actual performance. Specifically, absolute accuracy is operationalized as the difference between learners' judgments and performance regardless of the direction of the difference (i.e., $|X_{\text{Judgment}} - X_{\text{Performance}}|$). Positive and negative differences are both counted as inaccuracies and—in contrast to the bias measure—do not cancel each other out. Even though the absolute accuracy index does not provide information about the

Table 2 *Subjective Cognitive Load Questions for the Example Generation and Example Evaluation Concerning the Concept of Attribution (Translated from German).*

	Task Difficulty	Mental Effort
Example Generation	The difficulty of coming up with an example that illustrated the concept of attribution was…1: very low – 7: very high	My invested effort during coming up with an example that illustrated the concept of attribution was…1: very low – 7: very high
Example Evaluation	The difficulty of evaluating the quality of my example that illustrated the concept attribution was…1: very low – 7: very high	My invested effort during evaluating my example that illustrated the concept of attribution was…1: very low – 7: very high

direction of judgment accuracy, it clearly depicts the magnitude of accuracy from perfect accuracy (here: 0%) to perfect inaccuracy (here: 100%).

Third, we computed intra-individual *gamma correlations* (*G*) between the students' and the experts' ratings, indicating *relative accuracy*, which in this context refers to the students' ability to discriminate between high-, medium- and low-quality examples. Gamma correlations are the established measure in metacomprehension research for determining relative accuracy (e.g., De Bruin et al., 2011; Prinz et al., 2020; Thiede & Anderson, 2003). Correlations between -1 and 1 are possible, with higher positive values characterizing greater relative accuracy.

Assessment of Cognitive Load: Subjective Task Difficulty and Mental Effort

Subjective task difficulty and mental effort were measured in two phases of the experiment: during example generation and during evaluating the quality of the examples. These measures were not implemented in the study by Zamary et al. (2016). In terms of the example generation phase, after each generated example we asked the students to rate the task difficulty and the invested mental effort on 7-point Likert scales (1: *very low difficulty/mental effort*, 7: *very high difficulty/mental effort*). The wording of the questions can be found in Table 2 and was adapted from previously existing scales for the assessment of mental effort and perceived difficulty in task processing (see Paas, 1992, Paas et al., 2003; Schmeck et al., 2015). Concerning the evaluation phase, after each overall judgment (i.e., after assigning no, partial or full credit to an example), the learners were asked to rate task difficulty and invested mental effort as well. For the later analyses, the eight ratings in generating examples and the eight ratings in evaluating the examples were averaged for each of the two measures (.71 < Cronbach's $\alpha < .88$).

Procedure

The experiment was conducted in an online learning environment. All participants worked individually on their own devices. After the written informed consent was given, the participants were asked to fill out a demographic questionnaire and provide their grade point average (GPA). Next, the participants answered questions about their academic self-concept and took the pretest. Then, the expository text was presented with the instruction to read it carefully. Like in Zamary et al. (2016), the learners had four minutes to read the text and were then automatically forwarded to the next page. Subsequently, like in the study by Zamary et al., each concept definition was presented individually for self-paced study. After the initial learning phase, the participants were prompted to generate an example while the concept definition was visible during the whole example generation process. Immediately after generating the examples, the students answered questions concerning task difficulty and mental effort during generating the example (these questions were not part of the procedure in the study by Zamary et al.). In the next step, the participants were asked to evaluate the quality of their example. After making the judgment, the students were prompted to rate the task difficulty and mental effort in forming their judgment (these questions were not part of the procedure in the study by Zamary et al.). This procedure was repeated for each of the eight concepts. The experiment lasted approximately one hour.

Results

We used an α-level of .05 for all tests. As the effect size measure, we report Cohen's *d* for *t* tests and η_p^2 for *F* tests. Based on Cohen (1988), values around $d = 0.20$ and $\eta_p^2 = .01$ can be considered as small effects, values around $d = 0.50$ and $\eta_p^2 = .06$ as medium effects, and values around $d = 0.80$ or $\eta_p^2 = .14$ or higher as large effects. The mean scores and standard deviations for all groups on all measures are shown in Table 3. Data and analysis scripts can be viewed under https://doi.org/10.17605/OSF.IO/2BSP6.

Preliminary Analyses

In the first step, we tested whether the random assignment resulted in comparable groups. A 2×2-factorial ANOVA did not show any statistically significant effects regarding GPA, $F(1, 127) = 0.01$, $p = .913$, $\eta_p^2 < .01$ for expert example standards, $F(1, 127) = 0.15$, $p = .697$, $\eta_p^2 < .01$ for idea unit standards, and $F(1, 127) = 2.66$, $p = .105$, $\eta_p^2 = .02$ for the interaction effect. Similarly, the groups did not differ in terms of the academic self-concept, $F(1, 127) = 0.01$, $p = .933$, $\eta_p^2 < .01$ for expert example standards, $F(1, 127) = 0.02$, $p = .883$, $\eta_p^2 < .01$ for idea unit standards, and $F(1, 127) = 2.57$, $p = .111$, $\eta_p^2 = .02$ for the interaction effect, and prior knowledge, $F(1, 127) = 0.23$, $p = .631$, $\eta_p^2 < .01$ for expert example standards, $F(1, 127) = 0.06$, $p = .812$, $\eta_p^2 < .01$ for idea unit standards, and $F(1, 127) = 0.92$, $p = .338$, $\eta_p^2 < .01$ for the interaction between both factors. Jointly, these findings indicate that the random assignment resulted in comparable groups.

We also tested if there were any significant differences regarding the quality of the learner-generated examples, because when measures to foster judgment accuracy affect task performance, potential benefits regarding judgment accuracy can simply be due to the effects on performance. The ANOVA did not reveal any statistically significant main effects, $F(1, 127) = 0.82$, $p = .368$, $\eta_p^2 = .01$ for expert example standards, and $F(1, 127) = 0.86$, $p = .355$, $\eta_p^2 = .01$ for idea unit standards. The interaction effect also did not reach statistical significance, $F(1, 127) = 0.05$, $p = .818$, $\eta_p^2 < .01$. Hence, the effects of the support measures on judgment accuracy (see below) cannot be attributed to effects on task performance.

In view of recent findings which indicate that learners partly base their judgments of performance on the cognitive load they experience during the task (see Baars et al., 2020), which would render potential effects of the support measures on cognitive load during performing the task problematic, we also analyzed whether the two support measures affected

Table 3 Means (and Standard Deviations) of all Measures.

	Idea unit standards group ($n = 32$)	Expert example standards group ($n = 35$)	Expert example and idea unit standards group ($n = 35$)	Control group ($n = 29$)
Grade Point Average (min: 1; max: 6)	2.06 (0.59)	2.02 (0.71)	2.26 (0.67)	2.20 (0.78)
Prior Knowledge (min: 0; max: 100)	12.50 (7.10)	13.57 (10.22)	11.79 (8.68)	11.42 (7.73)
Academic Self-Concept (min: 1; max: 5)	3.61 (0.49)	3.60 (0.60)	3.46 (0.49)	3.43 (0.66)
Quality of Examples (min: 0; max: 100)	40.63 (18.38)	46.25 (17.43)	44.11 (17.35)	44.18 (16.86)
Task Difficulty (Generating Examples, min: 1; max: 7)	3.64 (1.24)	3.31 (0.99)	3.72 (1.01)	3.42 (0.97)
Mental Effort (Generating Examples, min: 1; max: 7)	3.90 (1.08)	3.60 (1.12)	4.01 (1.19)	3.96 (1.21)
Bias (min: -100; max: 100)	29.49 (20.28)	18.21 (18.77)	17.14 (13.92)	28.23 (20.50)
Bias for Concepts with 2 Idea Units (min: -100; max: 100)	35.94 (25.35)	20.71 (28.11)	22.14 (23.30)	38.79 (26.38)
Bias for Concepts with 3 Idea Units (min: -100; max: 100)	28.52 (21.12)	18.93 (21.08)	15.71 (19.96)	26.29 (20.14)
Bias for Concepts with 4 Idea Units (min: -100; max: 100)	25.00 (36.48)	14.29 (30.49)	15.00 (25.87)	21.55 (36.43)
Absolute Accuracy (min: 0; max: 100)	36.91 (16.60)	29.64 (11.78)	27.50 (11.86)	36.85 (14.60)
Absolute Accuracy for Concepts with 2 Idea Units (min: 0; max: 100)	37.50 (26.18)	33.57 (20.95)	26.43 (24.96)	42.24 (22.26)
Absolute Accuracy for Concepts with 3 Idea Units (min: 0; max: 100)	33.98 (18.85)	28.93 (13.81)	27.14 (15.60)	32.33 (16.88)
Absolute Accuracy for Concepts with 4 Idea Units (min: 0; max: 100)	42.19 (23.28)	27.14 (22.99)	29.29 (23.08)	40.52 (25.37)
Relative Accuracy (min: -1; max: 1)	0.37 (0.71)	0.59 (0.56)	0.55 (0.51)	0.34 (0.67)
Correlation between learners' ratings and word count (min: 0; max: 1)	0.21 (0.43)	0.23 (0.39)	0.28 (0.43)	0.11 (0.53)
Task Difficulty (Evaluating Examples, min: 1; max: 7)	2.86 (1.30)	3.09 (0.95)	3.44 (1.10)	2.89 (1.03)
Mental Effort (Evaluating Examples, min: 1; max: 7)	2.94 (1.27)	3.14 (1.05)	3.69 (1.18)	3.06 (1.22)

subjective task difficulty and mental effort invested in generating the examples. The support measures affected neither task difficulty, $F(1, 127) = 0.01$, $p = .944$, $\eta_p^2 < .01$ for expert example standards, $F(1, 127) = 2.85$, $p = .094$, $\eta_p^2 = .02$ for idea unit standards, and $F(1, 127) = 0.26$, $p = .612$, $\eta_p^2 < .01$ for the interaction, nor mental effort, $F(1, 127) = 0.38$, $p = .538$, $\eta_p^2 < .01$ for expert example standards, $F(1, 127) = 0.78$, $p = .380$, $\eta_p^2 = .01$ for idea unit standards, and $F(1, 127) = 1.38$, $p = .243$, $\eta_p^2 = .01$ for the interaction. Hence, effects on judgment accuracy cannot be explained via effects of the two measures of interest on cognitive load during example generation.

RQ 1: How Accurate Can Learners Evaluate the Quality of Self-Generated Examples? The students showed substantial inaccuracy in evaluating the quality of their examples. Overall, they were overconfident with a mean bias of 22.9% ($SD = 19.1\%$), which was significantly greater than zero, $t(130) = 13.75$, $p < .001$, $d = 1.20$. The overall absolute accuracy was 32.4% ($SD = 14.2\%$), which was also significantly greater than zero, $t(130) = 26.08$, $p < .001$, $d = 2.28$, indicating that the learners partly also underconfidently judged their examples.

Exploring learners' judgment accuracy more deeply, we computed bias and absolute accuracy scores for examples that received no, partial and full credit by the experts separately. Examples that received partial credit by the experts were most frequent with 54.3%, followed by no credit with 29.0% and full credit with 16.7% of the examples. We found a bias and absolute accuracy of 52.7% ($SD = 27.0\%$) for examples that received an expert rating of zero (bias and absolute accuracy are identical because it was not possible to underconfidently judge these examples), which significantly differed from zero, $t(108) = 20.33$, $p < .001$, $d = 1.95$. For examples that received partial credit, bias was 18.9% ($SD = 19.1\%$) and absolute accuracy was 27.3% ($SD = 13.4\%$), both of which were significantly greater than zero, $t(127) = 11.13$, $p < .001$, $d = 0.98$, and $t(127) = 22.95$, $p < .001$, $d = 2.02$. For correct examples (i.e., full credit in expert ratings), we found a bias of -16.8% ($SD = 21.4\%$) and an absolute accuracy of 16.8% ($SD = 21.4\%$; these numbers are identical in absolute terms because these examples could not be judged overconfidently), which significantly differed from zero, $t(89) = 7.46$, $p < .001$, $d = 0.79$, indicating that learners were underconfident in judging these examples.

Despite these findings that indicate substantial inaccuracy, there was evidence for some accuracy as well. The students rated correct examples significantly higher than examples that received partial credit, $t(89) = 5.60$, $p < .001$, $d = 0.59$. Also, partially correct examples were rated significantly higher than incorrect examples, $t(105) = 6.19$, $p < .001$, $d = 0.60$. Moreover, relative accuracy was significantly greater than zero, $G = .48$, $p < 001$. Jointly, these findings highlight that the students did show at least some accuracy in evaluating the relative quality of their generated examples. Fig. 2 provides an overview of the students' ratings as a function of the actual quality of the examples.

RQ 2: Does the Provision of Idea Unit Standards and Expert Example Standards Enhance Students' Judgment Accuracy in Evaluating the Quality of Self-Generated Examples? We were interested in whether the outlined (in)accuracy in learners' evaluation of their examples depended on whether learners received idea unit standards (RQ 2a) and expert example standards (RQ 2b). For explorative purposes, we were also interested in whether these two measures would interact regarding judgment accuracy. In terms of overall bias, a 2×2-factorial ANOVA revealed no statistically significant main effect for idea unit standards, $F(1, 127) < 0.01$, $p = .977$, $\eta_p^2 < .01$. By contrast, there was a statistically

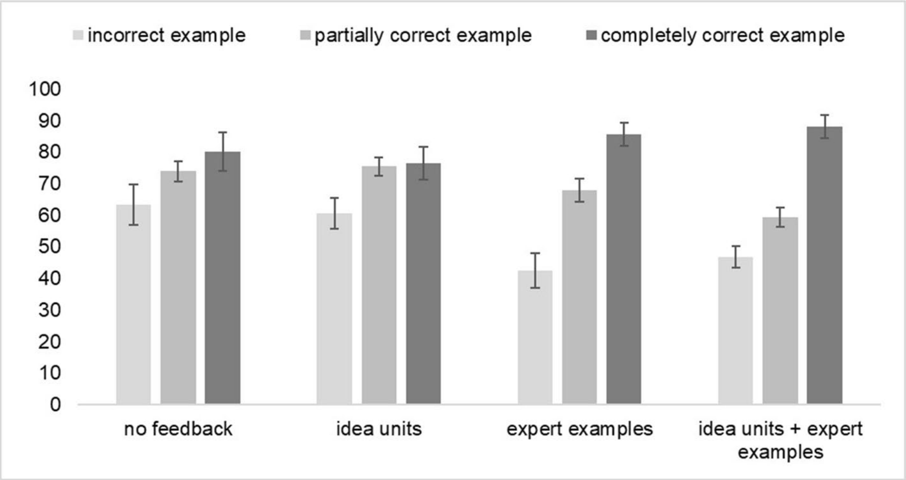


Fig. 2 Magnitude of students' judgments of the quality of their generated examples, as a function of the actual example quality (based on expert ratings) for each group. Error bars report standard error of the mean

significant main effect of expert example standards, $F(1, 127) = 12.00$, $p = .001$, $\eta_p^2 = .09$. Students who received expert example standards showed lower overall bias. No statistically significant interaction effect was found, $F(1, 127) = 0.13$, $p = .719$, $\eta_p^2 < .01$.

In terms of overall absolute accuracy, the pattern of results was similar. There was no statistically significant main effect of idea unit standards and no statistically significant interaction effect, $F(1, 127) = 0.19$, $p = .666$, $\eta_p^2 < .01$, and $F(1, 127) = 0.21$, $p = .648$, $\eta_p^2 < .01$. However, there was a statistically significant main effect of expert example standards, indicating higher absolute accuracy (i.e., values closer to zero) in the groups with expert example standards, $F(1, 127) = 11.90$, $p = .001$, $\eta_p^2 = .09$.

In terms of relative accuracy, an ANOVA did not yield a statistically significant effect for the factor expert example standards, $F(1, 119) = 3.65$, $p = .058$, $\eta_p^2 = .03$, nor for the factor idea unit standards or the interaction effect, $F(1, 119) < 0.01$, $p = .995$, $\eta_p^2 < .01$, and $F(1, 119) = 0.09$, $p = .769$, $\eta_p^2 < .01$, respectively.

In the next step, we analyzed the effects of the two support measures separately for incorrect, partially correct, and correct examples. For incorrect examples, concerning bias an ANOVA revealed a statistically significant main effect of expert example standards, $F(1, 105) = 11.89$, $p = .001$, $\eta_p^2 = .10$. The learners who received expert example standards showed lower bias. The effects of idea unit standards and the interaction did not reach statistical significance, $F(1, 105) = 0.03$, $p = .868$, $\eta_p^2 < .01$, and $F(1, 105) = 0.47$, $p = .495$, $\eta_p^2 < .01$ (for absolute accuracy, the results are identical because these examples did not allow underconfident judgments).

For partially correct examples, the pattern of results was slightly different. With respect to bias, as for the incorrect examples we found a statistically significant main effect of expert example standards, $F(1, 124) = 11.54$, $p = .001$, $\eta_p^2 = .09$, but no statistically significant effects for idea unit standards, $F(1, 124) = 1.19$, $p = .278$, $\eta_p^2 = .01$, or the interaction, $F(1, 124) = 2.40$, $p = .124$, $\eta_p^2 = .02$. Bias was lower when the students were provided with expert example standards. In terms of absolute accuracy, however, we did not find any statistically significant effects. Neither the main effect of idea unit standards, $F(1, 124) = 2.99$, $p = .087$, $\eta_p^2 = .02$, nor the main effect of expert example standards, $F(1, 124) = 3.20$,

p =.076, η_p^2 =.02, nor the interaction, F(1, 124) = 0.92, p =.339, η_p^2 < .01, were statistically significant.

For the correct examples, we again found a different pattern of results. Regarding both bias and absolute accuracy (note that these measures did not differ in absolute terms for these examples because learners could not overconfidently judge the correct examples), the ANOVA did not show any statistically significant effects, F(1, 86) = 0.02, p =.904, η_p^2 < .01 for idea unit standards, F(1, 86) = 3.44, p =.067, η_p^2 =.04 for expert example standards, and F(1, 86) = 0.45, p =.503, η_p^2 =.01 for the interaction.

Exploratory Analyses

To gain some exploratory insight into the cues that learners potentially utilized in evaluating their examples, we determined the gamma correlation between the length of learners' examples and their evaluations. We found a gamma correlation of G =.22, which indicates that learners tended to give higher judgments to lengthier examples and thus, at least in part, might have used the length of their examples as cue. Following up on this result, in the next step we analyzed whether the learner generated examples and the expert examples differed in length. For all eight declarative concepts, we found that the learner generated examples included fewer words than the expert examples, -39.65 < t(130) < -3.69, all ps < .001, -3.46 < d < -0.32. Jointly, these findings render it plausible that the better bias and absolute accuracy scores on part of the learners in the expert example groups, which were substantially driven by the learners lowering their self-evaluations, were in part due to the contrast regarding example length.

Analyzing the length differences furthermore revealed that the differences in length between the expert examples and learner generated examples depended on the number of idea units of which the concepts consisted (i.e., complexity of the concepts), F(2, 260) = 138.19, p < .001, η_p^2 =.52. Hence, in a second exploratory analysis we analyzed whether the effects of the expert example standards on bias and absolute accuracy depended on concept complexity. In the first step, we analyzed whether example quality and learners' evaluations depended on concept complexity. We found a statistically significant effect of complexity on example quality, F(2, 254) = 8.68, p < .001, η_p^2 =.06; the examples of the four-idea-unit concepts were of the highest quality, followed by the examples concerning the two-idea-unit and the examples concerning the three-idea-unit concepts. Also, learners' evaluations depended on concept complexity, F(2, 254) = 16.69, p < .001, η_p^2 =.12. The learners evaluated their examples of the two-idea-unit concepts the highest, followed by the examples regarding the four-idea-unit and the three-idea-unit concepts. In the second step, we addressed the potential dependency of the effects on bias and absolute accuracy on concept complexity. We found that bias depended on complexity; the concepts that included only two idea units yielded higher bias than the concepts that included three or four idea units, F(2, 254) = 7.40, p < .001, η_p^2 = .06. However, neither the effect of expert example standards on bias, F(2, 254) = 0.98, p = .378, η_p^2 = .01, nor the effect of idea unit standards on bias, F(2, 254) = 0.16, p = .843, η_p^2 < .01, depended on concept complexity. Regarding absolute accuracy, we did not find a statistically significant effect of concept complexity, F(2, 254) = 2.17, p = .116, η_p^2 = .02. Furthermore, like in terms of bias, neither the effect of expert example standards nor the effect of idea unit standards on absolute accuracy depended on concept complexity, F(2, 254) = 1.46, p = .235, η_p^2 = .01, and F(2,

254) = 1.49, p = .227, η_p^2 = .01, respectively. Jointly, these findings indicate that concept complexity did not matter for the benefits of expert example standards.

RQ 3: Does the Provision of Idea Unit Standards and Expert Example Standards Affect Task Difficulty and Mental Effort in Evaluating the Quality of Self-Generated Examples? In Research Question 3, we were interested in whether the idea unit standards and expert example standards would affect subjective task difficulty and invested mental effort in evaluating the quality of self-generated examples (RQ 3a and 3b, respectively). In terms of task difficulty, an ANOVA showed no statistically significant main effect of idea unit standards, F(1, 127) = 0.66, p = .419, η_p^2 = .01. Also, there was no statistically significant interaction effect, F(1, 127) = 1.00, p = .319, η_p^2 = .01. However, there was a statistically significant main effect of expert example standards, F(1, 127) = 4.16, p = .043, η_p^2 = .03. The learners who received expert example standards experienced evaluating their examples as more difficult than their counterparts.

Regarding mental effort, the pattern of results was similar. The ANOVA revealed a statistically significant main effect of expert example standards, F(1, 127) = 3.93, p = .049, η_p^2 = .03, indicating that the learners with expert example standards invested more mental effort than their counterparts. By contrast, no statistically significant effects were found for idea unit standards, F(1, 127) = 1.05, p = .308, η_p^2 = .01, or the interaction between both factors, F(1, 127) = 2.68, p = .104, η_p^2 = .02.

Discussion

The present study entails the following three main contributions. First, it can be concluded that providing learners with expert example standards substantially enhances accuracy in evaluating the quality of self-generated examples. Evidently, the learners who received expert example standards showed lower bias, as well as higher absolute accuracy than their counterparts without expert example standards. Second, providing learners with idea unit standards does not affect accuracy in evaluating the quality of self-generated examples. This finding conceptually replicates core findings of Zamary et al. (2016) and thus substantially strengthens the conclusion that idea unit standards are not helpful in the present context. Considered together, the two conclusions point to a third contribution of the present study, which is at the theoretical level. Specifically, our study indicates that, to effectively enhance judgment accuracy, standards, according to Rawson and Dunlosky's (2007) original idea, need to be designed such that they represent concrete correct responses to the respective tasks.

Why Expert Example Standards Did Help, but Idea Unit Standards Did not

Similar to Zamary et al. (2016), overall we found evidence for substantial inaccuracy in learners' evaluations of their examples (RQ 1). More specifically, in evaluating examples that received an expert rating of zero, the learners showed the largest inaccuracy, which likely reflects the *double curse of incompetence*, which was established by Dunning et al. (2003). When learners are not able to generate quality examples that at least partly correctly illustrate the respective concept's idea units, they most probably also lack the knowledge that is necessary to accurately evaluate the respective examples. Although to a substantially lower degree, however, the learners' evaluations were inaccurate for examples that received partial or full credit by the experts as well. Hence, even when the learners had the knowledge that was necessary to form quality examples, they did not or were not able to sufficiently use this knowledge to accurately evaluate their examples.

Nevertheless, in line with the study by Zamary et al. (2016) we also found evidence for a certain degree of accuracy in learners' self-evaluations. Overall, relative accuracy was well above zero, which indicates that the learners were fairly able to rank their judgments in terms of quality. Learners' judgment accuracy, however, depended on the type of support measures the learners received. More specifically, in terms of bias and absolute accuracy, the provision of expert example standards showed beneficial effects of medium size (RQ 2b). These benefits were largest in evaluating examples that received an expert rating of zero, and did not reach statistical significance for the examples that received full credit by the experts. In line with the above-mentioned notion that a low ability to generate quality examples would likely come along with low ability to accurately evaluate the examples, this finding indicates that the scaffolding by the expert example standards was most helpful when learners generated poor examples. For the evaluation of examples that received full credit by the experts, by contrast, the expert example standards might have been inappropriate. Potentially, during the evaluation of these quality examples, the learners might have benefitted from low and medium quality examples as external standards, for comparing them with their own examples would have highlighted both that the learners' examples were better and the threshold at which full credit should be assigned. Correspondingly, the additional provision of low- and medium-quality examples might also enhance relative accuracy, because it should help learners discriminate between partial and full credit examples in particular. Although the gamma correlations in the groups with expert example standards ($G = .55$ and $G = .59$) were substantially higher than the gamma correlations in the groups without expert example standards ($G = .34$ and $G = .37$) on the descriptive level, the expert example standards alone did not yield a statistically significant effect on relative accuracy in the present study (but note that the effect would have been significant if we had used one-tailed testing). Yet increases in relative accuracy would be highly desirable, because although the overall gamma correlation found in the present study ($G = .48$) was higher than the moderate value of $G = .24$ that was found in the recent meta-analysis on judgment accuracy in text comprehension by Prinz et al. (2020; see also Dunlosky & Lipko, 2007), it still means that more than 77% of the variance in actual example quality is not captured by learners' evaluations. Hence, efforts to further increase relative accuracy in future studies would be highly valuable.

As expected, the provision of expert example standards increased subjective task difficulty in the evaluation phase (RQ 3b). However, the processing of the expert example standards, which required the learners to read substantially more information than in the groups without expert example standards, and the transfer from the expert examples concerning what a good illustration would have to look like in the cover stories of the learners' examples obviously did not overload learners. Even in the group that received both types of standards and who thus had to process the most information in the evaluation phase, the average task difficulty rating was well below the mean of the scale (i.e., four) and the learners were still willing to overcome the increased task difficulty through increasing the mental effort they invested.

In terms of the effects of the idea unit standards, the pattern of results was less positive. Although we made slight adaptations to the materials, displays and procedure (e.g., by for example asking students to evaluate their cognitive load during example evaluation, see Method section) and hence our study should not be considered a direct but a conceptual replication of parts of the study by Zamary et al. (2016), our findings broadly support Zamary et al.'s notion that idea unit standards are not helpful in evaluating self-generated examples. Regarding none of the accuracy measures did we find any significant effects of the idea unit standards (RQ 2a). One explanation for this finding could be that other than the expert example standards, the idea unit standards did not represent concrete correct responses to the example generation tasks. Consequently, just like in generating the examples in the first place, in self-evaluating their examples learners had to generate an internal standard for what would constitute a correct illustration of each idea unit and then evaluate whether their examples would match the respective newly formed standards. In view of the finding that on average the objective example quality scarcely exceeded 40%, it is reasonable to assume that the forming of accurate internal standards on the basis of the idea unit standards was beyond learners' competence. The lack of effect of idea unit standards regarding subjective task difficulty and mental effort does not necessarily contradict this interpretation (RQ 3a). Rather, as it was far beyond learners' competence to form accurate internal standards based on the idea unit standards, the learners might have used the idea unit standards in a relatively shallow way and hence both task difficulty and invested mental effort were scarcely affected by the idea unit standards. Future studies that delve more deeply into learners' actual use of idea unit standards in evaluating self-generated examples (e.g., by think-aloud methodology) could test this tentative interpretation.

Jointly, these findings and conclusions point to one overarching theoretical contribution of the present study. To effectively enhance judgment accuracy, standards need to actually represent concrete correct responses to the respective tasks. To date, whenever standards represented concrete correct responses to the respective tasks (e.g., in retrieval practice tasks, see e.g., Rawson & Dunlosky, 2007; or in problem-solving tasks, see e.g., Baars et al., 2014), they largely enhanced judgment accuracy, which can be explained by the fact that in these conditions learners can compare their products to the standards in a relatively straightforward manner. Although concerning the thresholds when to assign partial or full credit to the respective products in particular such standards likely still leave some room for interpretation, they do not seem to overtax learners and hence result in increased judgment accuracy. By contrast, when the standards do not represent concrete correct responses to the respective task assignments and hence meaningful processing of the standards requires that learners would be able to correctly solve the respective task in the first place, standards are not beneficial. Although the latter notion needs to be challenged in future studies, in which standards that are not well aligned with the respective tasks are investigated in tasks other than example generation, it can tentatively be concluded that high consistency of the provided standards with concrete correct answers to the respective tasks is an essential design principle of effective standards.

Limitations and Future Research

It should be highlighted that the present study has some important limitations. First, we did not implement a posttest that assessed learners' comprehension of the eight declarative concepts after they generated the examples and had learners predict this performance. Consequently, we know neither whether the expert example (and idea unit) standards fostered metacomprehension such that learners were better able to accurately predict their performance on a subsequent posttest afterwards, nor whether the standards per se fostered comprehension. Also, we did not implement a subsequent learning phase and tested whether the standards fostered regulation in this phase, such that learners who received expert example (or idea unit) standards to a higher degree restudied the concepts that they had not yet fully understood. The benefits of the expert example standards on evaluating the quality of the learners' external products (i.e., the examples) that were found in the present study are nevertheless relevant to self-regulated learning, because learners likely frequently generate own examples and do not receive instructor feedback on their quality. In this case, the expert example standards could increase judgment accuracy, which can be expected to pave the way for effective regulation. Future studies that also assess the effects of the (expert example) standards on task performance, metacomprehension and regulation, however, are certainly needed to complement the present study's findings. Regarding effects on task performance, in research on standards in retrieval practice and problem-solving tasks, standards have been found to foster subsequent task performance (e.g., Baars et al., 2014; Rawson & Dunlosky, 2007). Hence, it could be expected that expert example standards would foster learners' ability to generate quality examples as well. In terms of metacomprehension and regulation, it could be assumed that expert example standards foster not only learners' ability to accurately evaluate their external products (the examples), but also help learners in forming diagnostic cues for evaluating their comprehension of the content (e.g., learners could use the degree to which they can explain why the expert examples would receive full credit as a cue). This, in turn, would render the expert example standards an actual *metacognitive* support measure. Based on the present study's findings, it cannot safely be said that the standards required any metacognitive processing in which the learners actually monitored their own understanding.

Second, the present study did not capture the specific processes that were executed by the learners in evaluating the quality of their examples. The findings of our exploratory analyses suggest that learners might have used example length as a cue, which, as the expert examples were on average substantially longer than the learners' examples, might have contributed to the better bias and absolute accuracy scores on part of the learners who received expert examples. Yet in the absence of measures that tap learners' actual processing of the expert examples (see Schalk et al., 2020), it remains unclear to what extent such relatively superficial comparisons actually contributed to the pattern of results and to what extent deep processing of the content of the expert examples mattered as well. Future studies that assess learners' cognitive and metacognitive processes in processing expert example standards and forming self-evaluations (for potential methodological approaches, see e.g., Dinsmore & Parkinson, 2013; Thiede et al., 2010; Van de Pol et al., 2020) are thus needed to appropriately understand how the expert example standards exert their beneficial effects. A related open issue concerning the process of evaluating the examples is whether the design of the scale on which learners evaluate the quality of their examples would matter. As, for replication purposes, the present study was closely aligned with the study of Zamary et al. (2016), we provided learners with a scale that consisted of only three quality levels (i.e., no, partial, or full credit). It is reasonable to assume, however, that providing learners with a more fine-grained scale might have helped them better discriminate between examples of different quality and reduce bias as well as increase absolute accuracy for examples that somehow fell between the categories. These benefits should be pronounced in particular when the scale on which learners evaluate the quality of their examples is closely aligned with the actual number of quality levels that can be found in learners' examples. Future studies should thus take a closer look at the role of the fit of the respective scales.

Third, it is important to highlight that the participants of our study had very low prior knowledge concerning the content domain and the declarative concepts that had to be illustrated. Accordingly, the average example quality was moderate at best (but comparable with the example quality reported in Zamary et al., 2016) and the learners likely had scarcely established any cues in evaluating examples concerning this content domain before taking part in our experiment. Consequently, the cues they utilized in self-evaluating their examples might have been relatively easy to change by the provision of expert example standards. Had the learners already been familiar with the content domain, by contrast, the effects might potentially have been lower since the cues that were suggested by the expert example standards might have competed with the learners' established cues. Hence, similar to findings in the field of example-based learning, which suggest that the benefits of providing examples decrease with increasing prior knowledge on part of the learners (e.g., Foster et al., 2018; Salden et al., 2010; for an overview, see Renkl, 2014), and more generally similar to research on the expertise reversal effect, which indicates that external guidance can be redundant or even interfere with the internal guidance of advanced learners (e.g., Chen et al., 2016; Kalyuga, 2007; Roelle & Berthold, 2013), the benefits of expert example standards might decrease with increasing prior knowledge. Future studies should therefore investigate whether expert example standards would have similar effects in content domains in which learners are familiar and already have significant prior knowledge, for this setting arguably would entail higher ecological validity than the setting used in the present study.

Acknowledgments We thank all students who participated in our study and the research assistant Chantal Cebula for assisting in collecting and analyzing the data. Furthermore, we thank Dennie Froese for his help with the computer programming.

Author's contributions Conceptualization: Linda Froese, Julian Roelle; Methodology: Linda Froese, Julian Roelle; Formal analysis and investigation: Linda Froese; Writing - original draft preparation: Linda Froese, Julian Roelle; Writing - review and editing: Julian Roelle; Supervision: Julian Roelle.

Funding Open Access funding enabled and organized by Projekt DEAL.

Code availability Not applicable

Declarations

Conflicts of Interests/Competing Interests The authors have no conflicts of interest to declare that are relevant to the content of this article.
Availability of data and material: Not applicable

Ethics approval All participants took part on a voluntary basis and gave written informed consent to their participation. All data were collected and analyzed anonymously. The study was conducted in full accordance with the German Psychological Society's (DGP's) ethical guidelines (2004, CIII; note that these are based on the APA's ethical standards) as well as the German Research Foundation's (DFG's) ethical standards. According to DFG, psychological studies only need approval from an institutional review board if a study exposes participants to risks that are related to high emotional or physical stress and/or if participants are not informed about the goals and procedures included in the study. As none of these conditions applied to the present study, we did not seek approval from an institutional review board.

Consent to participate Written informed consent was given for all students.

Consent for publication Not applicable

References

Baars, M., Vink, S., van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction*, *33*, 92–107. https://doi.org/10.1016/j.learninstruc.2014.04.004

Baars, M., Wijnia, L., de Bruin, A., & Paas, F. (2020). The relation between student's effort and monitoring judgments during learning: A meta-analysis. *Educational Psychology Review*, *32*(4), 979–1002. https://doi.org/10.1007/s10648-020-09569-3

Berthold, K., Nückles, M., & Renkl, A. (2007). Do learning protocols support learning strategies and outcomes? The role of cognitive and metacognitive prompts. *Learning and Instruction*, *17*(5), 564–577. https://doi.org/10.1016/j.learninstruc.2007.09.007

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

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review*, *30*(3), 703–725. https://doi.org/10.1007/s10648-018-9434-x

Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology*, *106*(3), 849–858. https://doi.org/10.1037/a0035934

Brod, G. (2021). Generative learning: Which strategies for what age? *Educational Psychology Review*, *33*(4), 1295–1318. https://doi.org/10.1007/s10648-020-09571-9

Byrne, B. M., & Shavelson, R. J. (1986). On the structure of adolescent self-concept. *Journal of Educational Psychology*, *78*(6), 474–481. https://doi.org/10.1037/0022-0663.78.6.474

Chen, O., Kalyuga, S., & Sweller, J. (2016). The expertise reversal effect is a variant of the more general element interactivity effect. *Educational Psychology Review*, *29*(2), 393–405. https://doi.org/10.1007/s10648-016-9359-1

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

De Bruin, A. B. H., Roelle, J., Carpenter, S., Baars, M., & EFG-MRE (2020). Synthesizing cognitive load and self-regulation theory: A theoretical framework and research agenda. *Educational Psychology Review, 32*(4), 903–915. https://doi.org/10.1007/s10648-020-09576-4

De Bruin, A. B. H., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology*, *109*(3), 294–310. https://doi.org/10.1016/j.jecp.2011.02.005

Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction*, *24*, 4–14. https://doi.org/10.1016/j.learninstruc.2012.06.001

Dunlosky, J., & Lipko, A. R. (2007) Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*(4), 228–232. https://doi.org/10.1111/j.1467-8721.2007.00509.x

Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *Quarterly Journal of Experimental Psychology*, *64*(3), 467–484. https://doi.org/10.1080/17470218.2010.502239

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*(3), 83–87. https://doi.org/10.1111/1467-8721.01235

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fiorella, L., & Mayer, R. E. (2013). The relative benefits of learning by teaching and teaching expectancy. *Contemporary Educational Psychology, 38,* 281–288. https://doi.org/10.1016/j.cedpsych.2013.06.001

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*(4), 717–741. https://doi.org/10.1007/s10648-015-9348-9

Foster, N. L., Rawson, K. A., & Dunlosky, J. (2018). Self-regulated learning of principle-based concepts: Do students prefer worked examples, faded examples, or problem solving? *Learning and Instruction*, *55*, 124–138. https://doi.org/10.1016/j.learninstruc.2017.10.002

Glogger, I., Schwonke, R., Holzäpfel, L., Nückles, M., & Renkl, A. (2012). Learning strategies assessed by journal writing: Prediction of learning outcomes by quantity, quality, and combinations of learning strategies. *Journal of Educational Psychology*, *104*(2), 452–468. https://doi.org/10.1037/a0026683

Gurung, R. A., Weidert, J., & Jeske, A. (2010). Focusing on how students study. *Journal of the Scholarship of Teaching and Learning, 10*(1), 28–35.

Hoogerheide, V., Visee, J., Lachner, A., & van Gog, T. (2019). Generating an instructional video as homework activity is both effective and enjoyable. *Learning and Instruction*, *64*, 101226. https://doi.org/10.1016/j.learninstruc.2019.101226

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

Kostons, D., van Gog, T., & Paas, F. (2009). How do I do? Investigating effects of expertise and performance-process records on self-assessment. *Applied Cognitive Psychology*, *23*(9), 1256–1265. https://doi.org/10.1002/acp.1528

Kostons, D., van Gog, T., & Paas, F. (2012). Training self-assessment and task-selection skills: A cognitive approach to improving self-regulated learning. *Learning and Instruction*, *22*(2), 121–132. https://doi.org/10.1016/j.learninstruc.2011.08.004

Lachner, A., Hoogerheide, V., van Gog, T. & Renkl, A. (2021). Learning-by-teaching without audience presence or interaction: When and why does it work? *Educational Psychology Review*. https://doi.org/10.1007/s10648-021-09643-4

Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using standards to improve middle school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied*, *15*(4), 307–318. https://doi.org/10.1037/a0017599

Lotz, C., Schneider, R., & Sparfeldt, J. R. (2018). Differential relevance of intelligence and motivation for grades and competence tests in mathematics. *Learning and Individual Differences*, *65*, 30–40. https://doi.org/10.1016/j.lindif.2018.03.005

Moning, J., & Roelle, J. (2021). Self-regulated learning by writing learning protocols: Do goal structures matter? *Learning and Instruction*, *75*, 101486. https://doi.org/10.1016/j.learninstruc.2021.101486

Nückles, M., Roelle, J., Glogger-Frey, I., Waldeyer, J., & Renkl, A. (2020). The self-regulation-view in writing-to-learn: Using journal writing to optimize cognitive load in self-regulated learning. *Educational Psychology Review, 32*(4), 1089–1126. https://doi.org/10.1007/s10648-020-09541-1

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

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*(1), 63–71. https://doi.org/10.1207/S15326985EP3801_8

Panadero, E., Brown, G. T., & Strijbos, J. W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational Psychology Review*, *28*(4), 803-830. https://doi.org/10.1007/s10648-015-9350-2

Prinz, A., Golke, S., & Wittwer, J. (2020). How accurately can learners discriminate their comprehension of texts? A comprehensive meta-analysis on relative metacomprehension accuracy and influencing factors. *Educational Research Review, 31,* 100358. https://doi.org/10.1016/j.edurev.2020.100358

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology*, *19*(4-5), 559–579. https://doi.org/10.1080/09541440701326022

Rawson, K. A., & Dunlosky, J. (2016). How effective is example generation for learning declarative concepts? *Educational Psychology Review*, *28*(3), 649–672. https://doi.org/10.1007/s10648-016-9377-z

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

Roelle, J., & Berthold, K. (2013). The expertise reversal effect in prompting focused processing of instructional explanations. *Instructional Science*, *41*(4), 635–656. https://doi.org/10.1007/s11251-012-9247-0

Roelle, J., & Berthold, K. (2017). Effects of incorporating retrieval into learning tasks: The complexity of the tasks matters. *Learning and Instruction*, *49*, 142–156. https://doi.org/10.1016/j.learninstruc.2017.01.008

Roelle, J., Berthold, K., & Fries, S. (2011). Effects of feedback on learning strategies in learning journals: Learner-expertise matters. *International Journal of Cyber Behavior, Psychology and Learning, 1,* 16–30. https://doi.org/10.4018/ijcbpl.2011040102

Roelle, J., Hiller, S., Berthold, K., & Rumann, S. (2017a). Example-based learning: The benefits of prompting organization before providing examples. *Learning and Instruction*, *49*, 1–12. https://doi.org/10.1016/j.learninstruc.2016.11.012

Roelle, J., Nowitzki, C., & Berthold, K. (2017b). Do cognitive and metacognitive processes set the stage for each other? *Learning and Instruction, 50,* 54–64. https://doi.org/10.1016/j.learninstruc.2016.11.009

Salden, R. J., Aleven, V., Schwonke, R., & Renkl, A. (2010). The expertise reversal effect and worked examples in tutored problem solving. *Instructional Science*, *38*(3), 289–307. https://doi.org/10.1007/s11251-009-9107-8

Schalk, L., Roelle, J., Saalbach, H., Berthold, K., Stern, E., & Renkl, A. (2020). Providing worked examples for learning multiple principles. *Applied Cognitive Psychology*, *34*(4), 813–824. https://doi.org/10.1002/acp.3653

Scheiter, K., Ackerman, R., & Hoogerheide, V. (2020). Looking at mental effort appraisals through a metacognitive lens: Are they biased? *Educational Psychology Review*, *32*(4), 1003–1027. https://doi.org/10.1007/s10648-020-09555-9

Schmeck, A., Opfermann, M., Van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: differences between immediate and delayed ratings. *Instructional Science*, *43*(1), 93–114. https://doi.org/10.1007/s11251-014-9328-3

Schnotz, W. (2010). Reanalyzing the expertise reversal effect. *Instructional Science*, *38*(3), 315–323. https://doi.org/10.1007/s11251-009-9104-y

Schöne, C., Dickhäuser, O., Spinath, B., & Stiensmeier-Pelster, J. (2002). *Skalen zur Erfassung des schulischen Selbstkonzepts: SESSKO*. Hogrefe.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, *4*(1), 33–45. https://doi.org/10.1007/s11409-008-9031-3

Simonsmeier, B. A., Flaig, M., Deiglmayr, A., Schalk, L., & Schneider, M. (2021). Domain-specific prior knowledge and learning: A meta-analysis. *Educational Psychologist*. https://doi.org/10.1080/00461520.2021.1939700

Steinmayr, R., & Spinath, B. (2009). The importance of motivation as a predictor of school achievement. *Learning and Individual Differences*, *19*(1), 80–90. https://doi.org/10.1016/j.lindif.2008.05.004

Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Thiede, K. W., & Anderson, M. C. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology*, *28*(2), 129–160. https://doi.org/10.1016/S0361-476X(02)00011-5

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes*, *47*(4), 331–362. https://doi.org/10.1080/01638530902959927

van de Pol, J., van Loon, M., van Gog, T., Braumann, S., & de Bruin, A. (2020). Mapping and drawing to improve students' and teachers' monitoring and regulation of students' learning from text: Current findings and future directions. *Educational Psychology Review, 32*(1), 951–977 https://doi.org/10.1007/s10648-020-09560-y

van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist*, *43*(1), 16–26. https://doi.org/10.1080/00461520701756248

Waldeyer, J., Heitmann, S., Moning, J., & Roelle, J. (2020). Can generative learning tasks be optimized by incorporation of retrieval practice? *Journal of Applied Research in Memory and Cognition*, *9*(3), 355–369. https://doi.org/10.1016/j.jarmac.2020.05.001

Waldeyer, J., & Roelle, J. (2021). The keyword effect: A conceptual replication, effects on bias, and an optimization. *Metacognition and Learning, 16*(1), 37–56. https://doi.org/10.1007/s11409-020-09235-7

Weinstein, C. E., & Mayer, R. E. (1986). The teaching of learning strategies. In M. C. Wittrock (Ed.), *Handbook on Research in Teaching* (3rd ed., pp. 315–327). MacMillan Reference Books.

Wylie, R., & Chi, M. T. (2014). The self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (2nd ed., pp. 413–432). Cambridge University Press.

Zamary, A., & Rawson, K. A. (2018). Which technique is most effective for learning declarative concepts—provided examples, generated examples, or both? *Educational Psychology Review*, *30*(1), 275–301. https://doi.org/10.1007/s10648-016-9396-9

Zamary, A., Rawson, K. A., & Dunlosky, J. (2016). How accurately can students evaluate the quality of self-generated examples of declarative concepts? Not well, and feedback does not help. *Learning and Instruction, 46*, 12–20 https://doi.org/10.1016/j.learninstruc.2016.08.002