Metacognition and Learning (2023) 18:923–944
https://doi.org/10.1007/s11409-023-09347-w

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

**Linda Froese**[1] · **Julian Roelle**[1]

Received: 23 September 2022 / Accepted: 28 May 2023 / Published online: 21 June 2023


**Abstract**
In acquiring new conceptual knowledge, learners often engage in the generation of examples that illustrate the to-be-learned principles and concepts. Learners are, however, bad at judging the quality of self-generated examples, which can result in suboptimal regulation decisions. A promising means to foster judgment accuracy in this context is providing external standards in form of expert examples after learners have generated own examples. Empirical evidence on this support measure, however, is scarce. Furthermore, it is unclear whether providing learners with poor examples, which include typical wrong illustrations, as negative example standards after they generated own examples would increase judgment accuracy as well. When they generated poor examples themselves, learners might realize similarities between their examples and the negative ones, which could result in more cautious and hence likely more accurate judgments concerning their own examples. Against this background, in a $2 \times 2$ factorial experiment we prompted $N = 128$ university students to generate examples that illustrate previously encountered concepts and self-evaluate these examples afterwards. During self-evaluation, we varied whether learners were provided with expert example standards (with vs. without) and negative example standards (with vs. without). In line with previous findings, expert example standards enhanced learners' judgment accuracy. The newly developed negative example standards showed inconsistent and partly even detrimental effects regarding judgment accuracy. The results substantiate the notion that expert example standards can serve as a promising means to foster accurate self-evaluations in example generation tasks, whereas negative example standards should be treated with caution.

**Keyword** Example generation; monitoring; judgment accuracy; standards

Learners are frequently engaged in generating own examples for previously studied new principles and concepts – either as a stand-alone activity or in the context of various established generative learning tasks such as journal writing (e.g., Nückles et al., 2020),

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

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

self-explaining (e.g., Wylie & Chi, 2014), or learning by teaching (e.g., Lachner et al., 2022). And that is a good thing – even when learners have only just studied definitions of new principles and concepts and have not received any further instruction on them, engaging learners in generating examples fosters learning outcomes (e.g., Rawson & Dunlosky, 2016; Roelle et al., 2022a). The benefits of generating examples on learning outcomes of course depend on the quality of the examples. Specifically, from both a theoretical and an empirical perspective, there is reason to believe that the effectiveness of example generation increases with increasing quality of the examples (e.g., Glogger et al., 2012; Rawson & Dunlosky, 2016; Roelle et al., 2022b). However, learners have substantial difficulties in accurately determining the quality of their examples (e.g., Froese & Roelle, 2022; Steininger et al., 2022; Zamary et al., 2016), which is problematic as it likely hinders learners' regulation decisions concerning when to continue or stop investing effort in example generation or when to seek help.

In view of this problem, which can reduce the effectiveness and efficiency of the common generative activity of example generation, in recent years there have been some attempts to design effective support measures that help learners to accurately evaluate self-generated examples (e.g., Froese & Roelle, 2022; Zamary et al., 2016). One support measure that proved to be particularly effective is providing learners with expert examples, which were explicitly highlighted as expert examples, as a comparison standard in judging their examples (Froese & Roelle, 2022). However, any finding should be viewed with some degree of scepticism until it is replicated. Furthermore, although *expert example standards* substantially increased judgment accuracy in the study by Froese and Roelle, learners were still largely inaccurate in evaluating examples of low quality. For instance, for examples that received 0% of the achievable points by expert raters, learners who received expert example standards gave themselves an average of approx. 42% of the points, which indicates that the provision of expert example standards alone cannot completely remedy learners' inaccuracy in evaluating self-generated examples.

Against this background, the present study's goals were twofold. First, we intended to replicate the benefits of expert example standards concerning judgment accuracy that were found by Froese and Roelle (2022). Second, we tested whether providing learners with typical poor examples, which were explicitly labelled as poor examples, as *negative example standards* would further reduce learners' overconfidence in evaluating self-generated examples, because learners might realize similarities between their poor examples and the negative ones, and hence serve as a beneficial add-on to expert example standards. To pursue these goals, we factorially varied whether learners were provided with expert example standards (with vs. without) and negative example standards (with vs. without) in evaluating self-generated examples. As main dependent variables, we determined the signed difference between students' and experts' quality ratings of the examples (i.e., bias), the absolute difference between these ratings regardless of the direction of the difference (i.e., absolute accuracy) as well as learners' ability to accurately rank their examples in terms of quality (i.e., relative accuracy).

## Example generation and evaluation: why it matters

The generation of examples implies that learners come up with concrete instances (e.g., real-world situations) that illustrate abstract content that is to be learned. For example, psychology students who are introduced to the concept of correspondence bias (i.e., 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) and asked to generate an own example for this new concept, could generate the example "Lisa receives a poor grade for her presentation. Her teacher explains this with the theory that Lisa does not have enough self-confidence to speak in front of the class. However, she forgets to consider that Lisa's grandmother died yesterday, which is why she could not concentrate during the presentation."

In terms of generative learning theory (e.g., Fiorella & Mayer, 2016), example generation can be conceived as a type of elaboration activity, which serves the function of integrating new learning content into existing knowledge structures and hence fosters comprehension of the respective content. Empirical studies support this theoretical notion. In a series of experiments, Rawson and Dunlosky (2016) showed that the generation of examples that illustrated newly acquired declarative concepts was more effective than restudying the concepts concerning learning outcomes (see also Roelle et al., 2022a). Notably, these benefits of generating examples occurred although learners only just read brief definitions of the declarative concepts beforehand and did not receive any further instruction such as provided examples or elaborate instructional explanations. Hence, although example quality and thus the effectiveness of generating examples can be expected to increase when learners already gained a certain degree of comprehension of the to-be-illustrated new content (e.g., Roelle et al., 2022a; Zamary et al., 2016), even in very early phases of acquiring new content generating examples can be beneficial. More often than as a stand-alone activity, generating examples arguably occurs in response to numerous generative learning tasks such as self-explanation tasks (e.g., Bisra et al., 2018), learning-by-teaching tasks (e.g., Lachner et al., 2022) or journal writing tasks (e.g., Moning & Roelle, 2021; Nückles et al., 2020). In the context of these tasks, example generation has been found to serve as a beneficial learning activity as well.

Since to date meaningful feedback on learner-generated examples can hardly be provided in an automatized manner, it is reasonable to assume that in deciding when to continue and when to stop investing effort in example generation, learners frequently must evaluate the quality of their examples on their own. However, several studies have shown that learners' performance in evaluating the quality of their examples is poor. For instance, Zamary et al. (2016) found that learners were largely overconfident. Specifically, the learners in their study awarded their examples with 75% on average (with 0% indicating lowest and 100% indicating highest example quality), whereas independent experts assessed the learners' examples with 36% on average. Hence, the learners overestimated the quality of their examples by nearly 40%. Since high judgment accuracy is crucial for effective regulation decisions (e.g., concerning when to invest further effort, when to seek help or when to stop investing effort in example generation) that foster learning outcomes and for efficient learning (e.g., Roelle et al., 2017; Thiede et al., 2003, 2017), this substantial inaccuracy is highly problematic.

In a conceptual replication, Froese and Roelle (2022) found similar deviations between experts' and learners' quality ratings. Examples that were assessed as incorrect by experts (i.e., scores of 0%) were overestimated by ca. 52%, examples rated by experts as partially correct (i.e., scores of 50%) reached a smaller but still significant overestimation by ca. 18%, and the evaluation of examples that were assessed as correct (i.e., scores of 100%) entailed some degree of underestimation (by ca. 16%). The massive overestimation of low-quality examples potentially mirrors parts of the *unskilled and unaware effect* that was introduced by Dunning et al. (2003). Learners who generate poor examples (and are therefore unskilled concerning the respective concepts that are to be illustrated) also lack the

knowledge to accurately judge example quality and hence are unaware of their examples' poor quality. This theoretical notion is in line with recent correlational findings showing that learner characteristics like low topic knowledge (i.e., unskilled) in combination with high self-perceived topic knowledge (i.e., unaware) reinforce overconfident judgments (see Golke et al., 2022). Beyond the replication of Zamary et al.'s findings concerning substantial overconfidence in evaluating self-generated examples, Froese and Roelle (2022) introduced a promising means to reduce the inaccuracy: the provision of *expert example standards*.

## Expert example standards: how they work and why they might be suboptimal when used alone

Standards are correct answers to a task assignment and designed for the explicit purpose of supporting learners in accurately judging the quality of self-generated products and solution steps (e.g., Baars et al., 2014; Lipko et al., 2009; Rawson & Dunlosky, 2007). And they do so effectively – in several tasks such as retrieval practice or problem-solving tasks, presenting the correct answer to the respective task as a comparison standard after a product or solution has been generated has shown to significantly enhance learners' judgment accuracy (e.g., Baars et al., 2014; Baker et al., 2010; Lipko et al., 2009; see also Waldeyer & Roelle, 2021). In tasks that require learners to generate own examples for previously introduced concepts, providing learners with *the* correct answers is hardly possible, because obviously a new concept can be correctly illustrated by many different examples that vary in their surface features (e.g., in their cover stories). Nevertheless, correct answers can serve as a beneficial comparison standard even in example generation tasks. In evaluating the quality of their own examples for newly acquired concepts, Froese and Roelle (2022) provided university students with two quality examples for each concept as a comparison standard after learners had generated their own examples. These *expert example standards*, which were provided in conjunction with the information that they would receive full credit when being scored, significantly fostered learners' accuracy in evaluating their own examples. Specifically, the expert example standards substantially reduced overconfidence in evaluating incorrect and partially incorrect examples ($\eta_p^2 = .10$ and $\eta_p^2 = .09$, respectively).

One explanation for these beneficial effects is that the expert example standards served as a beneficial basis for generating *cues* that were predictive for the actual quality of examples. According to *cue-utilization theory* (e.g., Koriat, 1997), in forming self-evaluations learners infer their judgments from various cues such as fluency during performing a task, the strategies used or previous task experiences (e.g., Baars et al., 2020; De Bruin et al., 2020; Thiede et al., 2010). Consequently, the accuracy of self-evaluations is theorized to depend on the degree to which the utilized cues are diagnostic (i.e., predictive) for the respective task performance (see also De Bruin et al., 2017). Comparing their examples with the expert ones potentially helped learners in monitoring cues of high diagnosticity and reduced the degree to which they based their judgments on hardly diagnostic cues such as example length or the perceived difficulty of generating the examples. This, in turn, resulted in the increase in judgment accuracy on part of the learners who received expert example standards that was observed by Froese and Roelle.

Although these findings should be viewed with a certain degree of skepticism until they are replicated, they suggest that expert example standards are a promising means to

enhance judgment accuracy in evaluating self-generated examples. However, it is important to highlight that the effects of expert example standards were nevertheless far from eliminating the problem of inaccurate judgments on part of the learners. Despite the provision of expert example standards, learners still overestimated their poor examples, which received no credit when scored by experts, by more than 40% (i.e., they credited more than 40% of the achievable points to their examples that received ratings of 0% by expert raters).

One explanation for this suboptimality of expert example standards could be that learners need a certain level of comprehension of the respective to-be-learned concept to be able to see why an expert example that illustrates the respective concept is of high quality. For instance, Hiller et al., (2020; see also Roelle & Renkl, 2020) found that learners who scarcely remembered previously explained principles and concepts were not able to self-explain subsequently provided examples. Hence, in using expert examples as a comparison standard, the aforementioned *unskilled and unaware effect* (Dunning et al., 2003) might apply as well. Generating poor examples (and therefore being unskilled in this context) likely reflects a low level of comprehension of the respective to-be-illustrated concept and leads not only to unawareness concerning the quality of the own examples, but also in terms of recognizing critical features of provided quality (i.e., expert) examples and using them to form diagnostic cues. That is, *unskilled* learners might not sufficiently be able to process the expert examples in a meaningful manner and hence see why their examples are of substantially lower quality than the expert ones. This, in turn, might result in overconfident judgments even when learners are provided with expert examples.

An instructional support measure that could help to remedy this suboptimality concerning self-evaluation of poor examples could be providing learners with typical poor examples, which are explicitly labelled as poor examples, as a type of *negative standards*. Specifically, when learners who generated poor examples process negative example standards, which include typical wrong illustrations and share further commonalities of poor examples such as omitted idea units that are not illustrated, they might realize similarities between their examples and the negative ones. This, in turn, might result in more cautious judgments concerning their own examples. Furthermore, such negative example standards might contribute to reducing the above-mentioned suboptimality of underconfidence concerning the self-evaluation of quality examples as well. Learners who have generated a correct example for a certain concept might see that their example differs substantially from the respective negative example, which might reduce their tendency to underconfidently judge their own examples.

Arguably, both potential beneficial effects of negative example standards on judgment accuracy can be expected to be higher when expert example standards are provided as well; that is, it is reasonable to assume that the effects of the two types of standards interact in terms of judgment accuracy. Research on contrasting cases and learning from comparisons indicates that contrasted examples can help learners discern the respective critical features (e.g., Gentner, 2010; Quilici & Mayer, 1996; Roelle & Berthold, 2015; Schalk et al., 2020). Hence, when both expert and negative example standards are provided, learners might be best able to directly compare both standard types and realize what constitutes poor and quality examples of the respective to-be-illustrated concepts. This, in turn, should enable learners to infer characteristics of both poor and quality examples and thus enhance the degree to which learners can generate predictive cues based on the example standards. These diagnostic cues should finally support learners in evaluating the quality of their own examples and thus increase judgment accuracy in evaluating self-generated examples. Combining expert and negative example standards could hence result in overadditive effects of the two types of standards (i.e., a beneficial interaction effect).

One caveat concerning the benefits of combining expert and negative example standards, however, could be that processing these standards and incorporating the generated cues in their judgments overloads learners. In their study on the effects of expert example standards, Froese and Roelle (2022) found that expert example standards increased perceived task difficulty and mental effort in self-evaluating self-generated examples, which however did not appear to detrimentally affect the benefits of expert example standards. Given that processing negative example standards would contribute to increases in cognitive load as well, the pattern of results might potentially change, because requiring learners to process both types of standards might be overtaxing. In this case, combining both types of standards might even have detrimental effects in comparison to providing either one of the standards alone, which would reflect a subadditive effect of combining the two types of standards (i.e., a detrimental interaction effect).

## The present study

In the present study, we pursued two main goals. First, we aimed at replicating the findings of Froese and Roelle (2022) concerning the effectiveness of expert example standards. We hypothesized that expert example standards would increase judgment accuracy in evaluating self-generated examples (Hypothesis 1a). Second, in view of the outlined theoretical considerations, we assumed that negative example standards would foster judgment accuracy in evaluating self-generated examples as well (Hypothesis 1b). Third, we were interested in potential interaction effects of the two support measures. Based on our theoretical considerations, which yielded that combining expert and negative example standards could result either in beneficial or detrimental interaction effects, we did not predict a specific interaction pattern but only hypothesized that expert and negative example standards would result in interaction effects concerning judgment accuracy (Hypothesis 1c). We were also interested in the effects of the two support measures on cognitive load during self-assessment. We assumed that both expert example standards and negative example standards would increase the cognitive load that learners experience in evaluating their examples (Hypothesis 2a and 2b).

## Method

### Sample and design

In view of the medium-sized effects of expert example standards that were reported in the study by Froese and Roelle (2022; i.e., $\eta_p^2 = .09$ regarding effects on absolute accuracy and bias) and the absence of previous studies on negative example standards, we used the lower bound of a medium effect size (i.e., $\eta_p^2 = .06$, see Cohen, 1988) as the basis for our a priori power analysis (further parameters: $\alpha = .05$, $\beta = .20$). The power analysis was conducted with G*Power 3.1.9.3 (Faul et al., 2007) and a $2 \times 2$ ANOVA, which corresponded with the present 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 $N = 128$ participants. Against this background, we recruited a sample of $N = 128$ students from various universities in Germany (95 female, 33 male; $M_{Age} = 25.16$ years, $SD_{Age} = 4.26$ years).

All students gave their written informed consent and received 10 € for their voluntary participation.

The study was designed in part as a replication of the study by Froese and Roelle (2022), which is why the procedure of the present study was closely aligned to Froese and Roelle's study. All learners read an expository text that covered eight concepts related to social cognition and behavioral patterns and subsequently studied definitions of the eight concepts. After this initial study phase, they were prompted to generate one example per concept while the concept definition was presented. Following a 2×2 factorial between-subject design, the learners then did or did not receive *expert example standards* and did or did not receive *negative example standards* while they were asked to assess their self-generated examples with no, partial or full credit. All participants were randomly assigned to one of the four conditions.

## Materials

### Expository text and concept definitions

In the initial study phase, all learners were asked to read an expository text of 476 words describing eight phenomena that related to human cognition and behavior in social situations (but without any concrete examples). Specifically, the text covered attribution theory, which describes that people seek causal explanations for the behavior of other people (concept of attribution). In the context of these attribution processes, the text differentiates between internally and externally motivated behavior, while latter is described as based on social conventions (concept of social norms). In the further course, the text addresses overlaps in behavior between one and the same person (e.g., concept of consistency), but also between other persons (concept of consensus). The last section deals with distorted perceptions of one's own and other people's behavior (concept of self-serving bias), of other people's behavior (concept of correspondence bias), and of the world (concept of just-world hypothesis). The text was identical to the one used in Froese and Roelle (2022), which, in turn, was based on material designed by Zamary et al. (2016). After the learners had finished reading the text, the concept definitions of the respective social phenomena (e.g., the concept of attribution was defined as *the process through which we seek to determine the causes behind people's behavior*) were then provided one after another for self-paced study. The learning material can be viewed under: https://doi.org/10.17605/OSF.IO/K3EMF.

### Support measures in evaluating the examples

Like in Froese and Roelle (2022, see also Zamary et al., 2016), all learners were prompted to generate an example for each concept after they had carefully studied the expository text and the concept definitions. The learners could reread the concept definitions throughout the entire generation process (i.e., open-book generative task; see Waldeyer et al., 2020). Immediately after each generation trial, the learners were asked to evaluate their example. The instruction to assess the self-generated examples ("If the quality of your example was being graded, what do you think it would receive?") as well as the rating scale (no credit, partial credit, or full credit) was the same in all conditions (adapted from Zamary et al. (2016) and identical to Froese and Roelle (2022)).

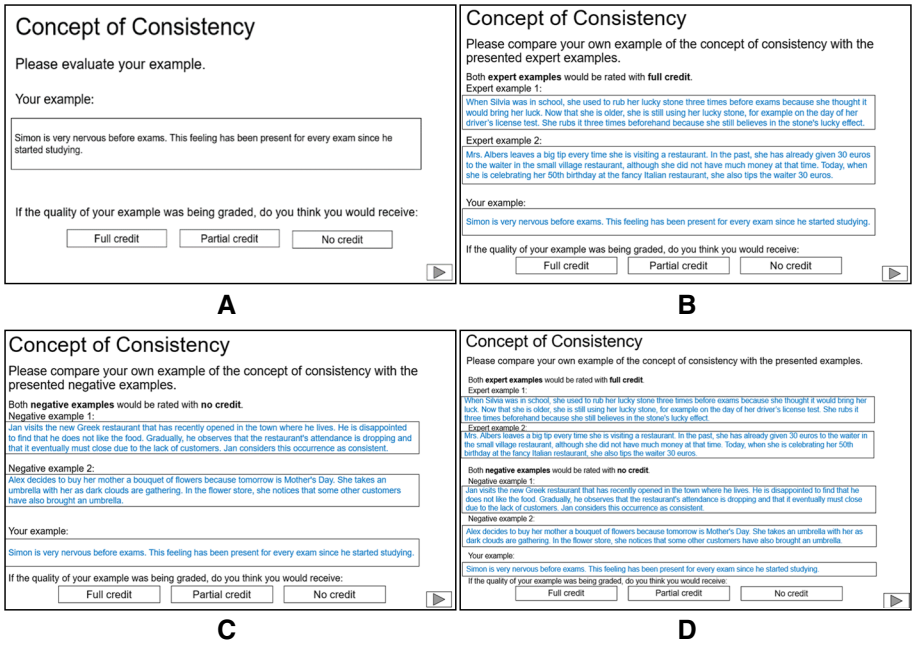


**Fig. 1** Screens of example evaluation for the concept of consistency for each group. **A** No Standards, **B** Expert Example Standards, **C** Negative Example Standards and **D** Combination of Expert and Negative Example Standards

During the self-assessment of the example quality, the learners in the control condition, who received neither expert example standards nor negative example standards, were shown only their example, the evaluation prompt and the rating scale asking the students to evaluate their example by assigning no, partial, or full credit. The set-up for these learners is illustrated in Panel A of Fig. 1.

The learners in the expert example standards condition received two expert examples per concept. To help learners identify the critical features of quality examples and prevent them from focusing on surface features, the two examples used different cover stories (i.e., *structure-emphasizing examples*, see Renkl, 2014). The two expert examples were provided on the screen above their self-generated example with the prompt to compare their example to the expert examples. The learners were informed that both expert examples would receive full credit (see Panel B in Fig. 1). The expert examples were the same as in Froese and Roelle (2022). In their study, Froese and Roelle developed and piloted these expert examples with the help of four experts who were well versed with the concepts. The experts rated each example independently and only examples that were assigned full credit by all experts were used (16 expert examples in total, i.e., two per concept).

In the negative example standards condition, the learners were shown two poor examples above their self-generated example and were informed that both negative examples would receive no credit, before asking them to judge the quality of their own example (see Panel C in Fig. 1). Like the expert examples, the poor examples used different cover stories to prevent learners from focusing on this surface feature which is hardly

predictive for example quality. The design of the negative examples was inspired by no credit examples from the study by Froese and Roelle (2022). Specifically, based on these no credit examples, typical errors in illustrating the eight concepts were identified and integrated in the negative example standards. For instance, when the learners in the previous study by Froese and Roelle tried to illustrate the concept of correspondence bias that was defined as *the tendency to attribute other people's behavior to internal causes to a greater extent than it is actually justified while underestimating the effect of the situation*, they often created examples that mixed up the role of internal and external causes. This as well as further errors were then integrated into the negative example standards. All 16 negative examples were examined by experts who assured that only examples that illustrated none of the components of the concepts correctly (i.e., no credit examples) were used as negative examples in the present study.

In the expert and negative example standards condition, two expert and two negative examples were provided above the own example with the respective information on how these four examples were being rated (expert examples with full credit, negative examples with no credit). The provision of four examples in total (two examples per standard type) was necessary, because otherwise the above-mentioned *structure-emphasizing function*, which was implemented to help learners identify the critical features of the respective examples, would have been lost. At the bottom of the slide, learners were shown the prompt to evaluate the self-generated example in comparison to the four provided examples (see Panel D in Fig. 1). All expert and negative examples can be viewed under https://doi.org/10.17605/OSF.IO/K3EMF.

## Instruments and measures

### Assessment of academic self-concept

We measured learners' academic self-concept as a control variable by means of an adapted version of a subscale of the SESSKO (Schöne et al., 2002), because the academic self-concept can be a strong motivational predictor of learners' performance. The five items of the subscale assessed learners' *absolute* self-concept (i.e., without any reference standard) on 5-point Likert scales, with higher scores indicating higher absolute self-concept (e.g., "*I am…* 1: *not intelligent* – 5: *very intelligent*"). The five ratings were aggregated for the later analyses (Cronbach's $\alpha = .84$).

### Pretest

Because prior knowledge is an important cognitive learning prerequisite (e.g., Simonsmeier et al., 2022), we measured learners' prior knowledge on the topic of the social phenomena with a pretest consisting of eight questions. The questions prompted the students to define the eight declarative concepts one after another. The answers were scored by two independent raters who evaluated if the respective idea units of the concepts were or were not included in the learners' answers. We computed intraclass correlation coefficients with measures of absolute agreement to determine the interrater reliability for each question. All ICCs were greater than .85. The scores were then summed up, averaged over all eight concepts, and converted into percentages (i.e., 0 – 100%, Cronbach's $\alpha = .49$).

### Assessment of example quality

To determine the quality of the learner generated examples, the concept definitions were separated into two, three or four idea units. For example, the concept of correspondence bias was split into the four idea units (1) *tendency to attribute other people's behavior*, (2) *attribution is misdirected*, (3) *overestimating the role of internal causes* and (4) *underestimating the role of external causes.* Two independent expert raters then assessed for each example if the respective idea units were or were not illustrated in the example. These ratings were then converted into no, partial, or full credit. For instance, for the concept of correspondence bias, one participant of the present study came up with the example "At the bakery, a woman pushes her way in. I think she does that because she is just a very rude person". This example was rated with partial credit as it entailed the idea units (1) *tendency to attribute other people's behavior* and (3) *overestimating the role of internal causes*. By contrast, the idea unit that this is a wrong interpretation (i.e., idea unit (2) *attribution is misdirected*) and the role of external causes (i.e., idea unit (4) *underestimating the role of external causes*) were not illustrated in the respective example. The interrater reliability was very high for all eight concepts (all ICCs > .85). The ratings were then summed up and averaged across all concepts, resulting in one example quality score for each learner (in form of percentages, i.e., 0 – 100%).

### Judgment accuracy

To examine learners' accuracy in evaluating their self-generated examples, we computed the three measures *bias*, *absolute accuracy,* and *relative accuracy* (see Schraw, 2009). For bias, which indicates the signed discrepancy between a judgment and the actual performance, we subtracted the experts' ratings from the students' ratings ($X_{\text{students ' judgment}} - X_{\text{experts ' judgment}}$), resulting in values between -100 and 100%. Negative values indicate underconfidence, values close to zero can be interpreted as high accuracy, and positive values indicate overconfidence.

Even though bias enables conclusions of the direction of potential inaccuracies (i.e., under- or overconfidence) it is not mindful of the possibility that over- and underconfident judgments may nullify each other within persons and groups, which is why we also calculated absolute accuracy. Absolute accuracy is determined as the absolute discrepancy between learners' and experts' ratings (i.e., $|X_{\text{students ' judgment}} - X_{\text{experts ' judgment}}|$) and allows conclusions about the correspondence of a self-evaluation and the actual performance without any annulments (but also without information on the direction of the deviation). Hence, values between 0 and 100% were possible, with values closer to zero indicating higher accuracy.

In the last step, we computed the intra-individual gamma correlations between the learners' and the experts' ratings, indicating relative accuracy. This measure of judgment accuracy does not take the absolute discrepancy between the judgment scores and the actual performance into account. It rather provides information on learners' ability to assess the relation between the quality levels of their examples, that is, the extent to which the examples are correctly ranked from low to high quality. The gamma correlations range from -1 to 1, where -1 indicates that the example quality was ranked inversely and 1 indicates that the ranks of the judgments correspond exactly to the actual quality ranking.

### Assessment of cognitive load

We measured learners' cognitive load during the generation of the examples (as a control variable) as well as during the example evaluation process (as a dependent variable). For this purpose, we used two items that captured learners' perceived task difficulty and mental effort (adapted from Paas, 1992; see also Schmeck et al., 2015). Both items were rated on 7-point Likert scales and adjusted for each task. Specifically, for the ratings of the perceived task difficulty and mental effort during the example generation, we asked: *The difficulty of/My invested effort during coming up with an example that illustrated the concept of attribution was…* 1: *very low*, 7: *very high*. To assess the task difficulty and mental effort during the example evaluation, we asked: *The difficulty of/My mental effort during evaluating the quality of my example that illustrated the concept of attribution was…* 1: *very low*, 7: *very high*, respectively. For each measure, we averaged the scores over all eight concepts (.71 < Cronbach's α < .88).

### Procedure

The learners passed the experiment in an online learning environment. First, they gave their written informed consent and filled out a questionnaire on personal information and on their academic self-concept. Subsequently, all learners took the pretest (self-paced). Afterwards, they read the expository text within a time frame of four minutes before being automatically proceeded to the next study phase. Here, all participants were asked to study one concept definition after another at their own pace. Next, the learners were asked to generate an example for the first concept in an open book format (i.e., the definition of the respective concept was presented during the task). Afterwards, all learners rated the task difficulty and the mental effort they experienced during the example generation task. In the next step, the experimental manipulation took place. The learners were asked to assess the quality of their self-generated example with no, partial or full credit while comparing their example to two expert examples, two negative examples, all four example standards or no comparison standard. The subjective task difficulty and mental effort during the example evaluation process were then assessed by all participants. The learners repeated this sequence (i.e., example generation, assessment of cognitive load, example evaluation, and assessment of cognitive load) for each concept until all learners generated and evaluated eight examples. The procedure took approx. one hour.

## Results

An α-level of .05 was used for all analyses. We report $\eta_p^2$ as effect size measure. Following Cohen (1988), values around $\eta_p^2 = .01$ indicate small effects, values around $\eta_p^2 = .06$ medium effects, and values around $\eta_p^2 = .14$ or higher large effects. The descriptive statistics (i.e., means and standard deviations) for all groups are shown in Table 1. Figure 2 provides an overview of the students' ratings as a function of the actual quality of the examples. Data and analysis scripts can be found under https://doi.org/10.17605/OSF.IO/K3EMF.

**Table 1** Means (and Standard Deviations) of all measures

| | Expert example standard group ($n=34$) | Negative example standard group ($n=30$) | Expert and negative example standard group ($n=34$) | Control group ($n=29$) |
|---|---|---|---|---|
| Prior Knowledge (0 – 100%) | 14.71 (10.07) | 10.21 (9.06) | 12.68 (8.77) | 6.47 (5.66) |
| Academic Self-Concept (1 – 5) | 3.44 (0.63) | 3.44 (0.63) | 3.56 (0.60) | 3.40 (0.68) |
| Quality of Examples (0 – 100%) | 42.46 (17.25) | 41.25 (14.91) | 44.30 (14.86) | 37.07 (16.69) |
| Task Difficulty (Generating Examples, 1 – 7) | 3.40 (1.11) | 3.51 (1.09) | 3.47 (1.01) | 3.53 (1.07) |
| Mental Effort (Generating Examples, 1 – 7) | 3.96 (1.26) | 3.93 (1.03) | 4.17 (1.23) | 3.75 (1.12) |
| Bias (-100 – 100%) | 22.43 (17.21) | 22.29 (21.56) | 23.16 (19.76) | 36.21 (23.88) |
| Absolute Accuracy (0 – 100%) | 33.09 (12.64) | 38.96 (13.79) | 33.82 (15.62) | 43.10 (16.31) |
| Relative Accuracy (-1 – 1) | 0.48 (0.57) | -0.02 (0.67) | 0.36 (0.63) | 0.36 (0.67) |
| Task Difficulty (Evaluating Examples, 1 – 7) | 3.25 (1.03) | 3.46 (1.04) | 3.33 (0.88) | 2.61 (0.98) |
| Mental Effort (Evaluating Examples, 1 – 7) | 3.51 (1.24) | 3.50 (1.14) | 3.77 (1.13) | 2.65 (1.15) |

Note that all measures depicted in this table were calculated with $N=127$ participants as one outlier was excluded

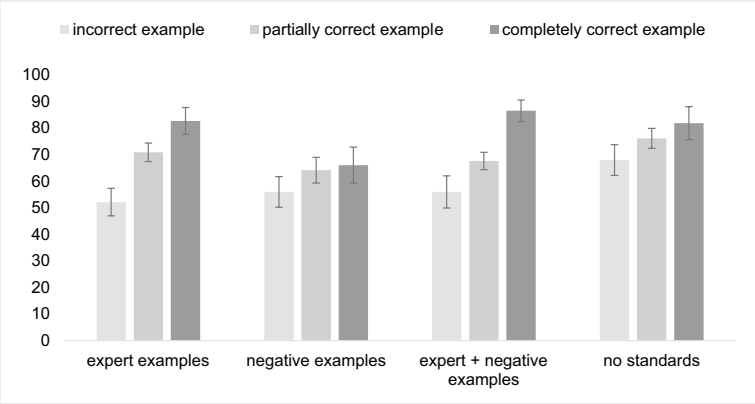


**Fig. 2** Magnitude of students' ratings of the quality of their self-generated examples as a function of the actual example quality (based on experts' ratings) for each group. Error bars report standard error of the means

## Preliminary analyses

First, we tested whether the random assignment resulted in comparable groups. Regarding academic self-concept, we did not find a statistically significant effect of condition, $F(3, 124) = 0.39$, $p = .759$, $\eta_p^2 < .01$. However, there was a statistically significant effect of condition concerning prior knowledge, $F(3, 124) = 4.37$, $p = .006$, $\eta_p^2 = .09$. The learners who received neither expert nor negative example standards appeared to have lower prior knowledge than the learners in the other three groups. Due to the random assignment of the participants to the conditions, this significant effect can be attributed to chance. It therefore can be controlled for by including prior knowledge as a covariate in our subsequent analyses (see Miller & Chapman, 2001). The assumption of homogeneous regression slopes was not violated in any of the analyses concerning our main dependent variables. Note that one outlier in the group that received negative example standards with a prior knowledge score of 50%, which was more than three standard deviations above the mean, did not substantially contribute to this pattern of results. In all subsequent analyses, we nevertheless excluded this participant and highlighted when the results of an analysis would substantially change when this outlier was not excluded.

## Effects of expert and negative example standards on judgment accuracy

We hypothesized that expert example standards (Hypothesis 1a) and negative example standards (Hypothesis 1b) would enhance the accuracy of learners' self-evaluations and that the two measures would interact concerning judgment accuracy (Hypothesis 1c). In terms of bias, only Hypothesis 1a was confirmed. We found a statistically significant main effect of expert example standards, $F(1, 122) = 5.35$, $p = .022$, $\eta_p^2 = .04$, indicating that the learners who received expert example standards showed lower bias than the learners without expert example standards. There was no significant main effect of negative example standards, $F(1, 122) = 3.69$, $p = .057$, $\eta_p^2 = .03$. Note that if the above mentioned outlier concerning prior knowledge was not excluded, the main effect of negative example standards would become statistically significant, $F(1, 123) = 4.10$, $p = .045$, $\eta_p^2 = .03$, which

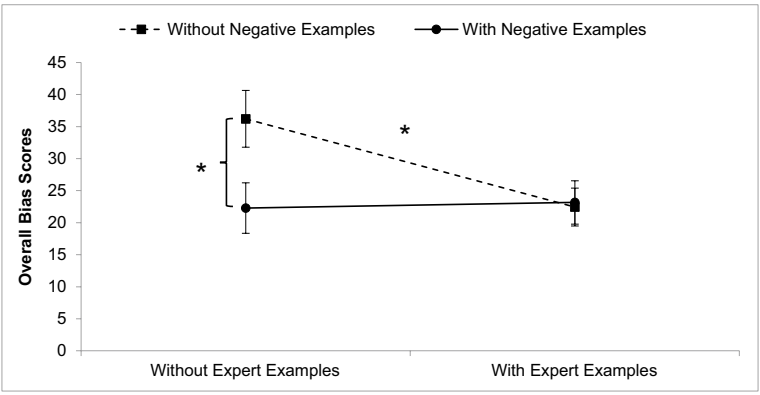


**Fig. 3** Interaction effect between expert and negative example standards regarding bias scores. Asterisks (*) indicate significant effects at $p < .05$. Error bars represent standard errors of the means

would indicate that the learners who received negative example standards showed lower bias than their counterparts. We also found a statistically significant interaction effect, $F(1, 122) = 5.46$, $p = .044$, $\eta_p^2 = .03$. The pattern of the interaction effect is shown in Fig. 3. Exploring the interaction revealed that expert example standards reduced bias for learners who did not receive negative example standards ($p = .010$, $\eta_p^2 = .10$), but not for learners who received negative example standards ($p = .895$, $\eta_p^2 < .01$). Similarly, negative example standards reduced bias for learners who did not also receive expert example standards ($p = .010$, $\eta_p^2 = .11$), but not for learners who did receive expert example standards ($p = .752$, $\eta_p^2 < .01$).

In terms of absolute accuracy, we found a different pattern of results. Like for bias and confirming Hypothesis 1a, we found a statistically significant main effect of expert example standards, $F(1, 122) = 11.42$, $p = .001$, $\eta_p^2 = .08$. The learners with expert example standards reached values closer to zero, showing better absolute accuracy than their counterparts without expert examples. By contrast, there was no statistically significant main effect of negative example standards, $F(1, 122) = 0.57$, $p = .450$, $\eta_p^2 < .01$, and also no statistically significant interaction effect, $F(1, 122) = 1.57$, $p = .213$, $\eta_p^2 = .01$.

Concerning relative accuracy, the pattern of results was again different. In line with Hypothesis 1a, we found that expert example standards enhanced relative accuracy, $F(1, 114) = 4.82$, $p = .030$, $\eta_p^2 = .04$.[1] We also found a statistically significant main effect of negative example standards, $F(1, 114) = 4.45$, $p = .037$, $\eta_p^2 = .03$. Surprisingly, contrary to Hypothesis 1b, negative example standards decreased rather than increased relative accuracy. Concerning Hypothesis 1c, there was no statistically significant interaction effect between the two factors, $F(1, 114) = 1.08$, $p = .300$, $\eta_p^2 < .01$.

Although the expert and negative example standards were provided after example generation, we nevertheless investigated if there are any unexpected effects of the standards regarding example quality (e.g., forward effects that transfer to the respective next examples because the comparison with the respective standards could potentially help learners

[1] Please note that $n = 8$ participants could not be included in these analyses due to indeterminate gamma correlations.

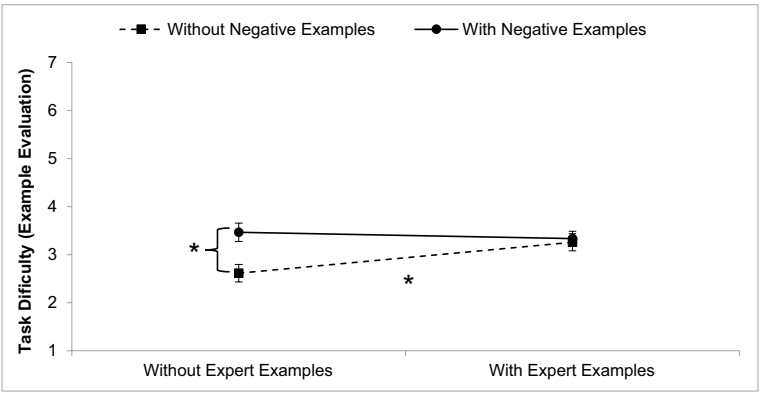


**Fig. 4** Interaction effect between expert and negative example standards regarding learners' perceived task difficulty during example evaluation. Asterisks (*) indicate significant effects at $p < .05$. Error bars represent standard errors of the means

see that they need to invest more effort in generating the next examples). We did not find any statistically significant differences between the groups concerning example quality, $F(1, 122) = 1.62$, $p = .205$, $\eta_p^2 = .01$ for expert example standards, $F(1, 122) = 1.06$, $p = .305$, $\eta_p^2 < .01$ for negative example standards, and $F(1, 122) = 0.10$, $p = .743$, $\eta_p^2 < .01$ for the interaction between both factors. Also, we tested whether there were any statistically significant effects concerning cognitive load during example generation, which is important because experienced cognitive load can be used as a cue to judge one's level of performance (see Baars et al., 2020). However, neither in terms of task difficulty (all $F$s $< 1$, all $p$s $> .850$, all $\eta_p^2 < .01$) nor in terms of mental effort (all $F$s $< 1.15$, all $p$s $> .285$, all $\eta_p^2 < .01$) did we find any significant main effects or interactions.

## Effects of expert and negative example standards on cognitive load during example evaluation

We hypothesized that the provision of expert example standards (Hypothesis 2a) and negative example standards (Hypothesis 2b) would increase learners' subjective cognitive load during example evaluation. Regarding task difficulty, we found a statistically significant main effect of negative example standards, $F(1, 122) = 7.15$, $p = .009$, $\eta_p^2 = .05$, but not of expert example standards, $F(1, 122) = 2.64$, $p = .107$, $\eta_p^2 = .02$. The negative example standards increased task difficulty. There was also a statistically significant interaction effect, $F(1, 122) = 5.30$, $p = .023$, $\eta_p^2 = .04$. The pattern of the interaction is depicted in Fig. 4. Exploring the interaction effect revealed that expert example standards enhanced task difficulty when learners did not also receive negative example standards ($p = .017$, $\eta_p^2 = .09$), but not when learners did receive negative example standards ($p = .656$, $\eta_p^2 < .01$). Similarly, negative example standards enhanced task difficulty when learners did not also receive expert example standards ($p = .002$, $\eta_p^2 = .15$), but not when learners did receive expert example standards ($p = .802$, $\eta_p^2 < .01$).

Concerning mental effort, we found that both types of standards increased mental effort, $F(1, 122) = 5.98$, $p = .016$, $\eta_p^2 = .04$ for expert example standards, and $F(1, 122) = 6.81$, $p = .010$, $\eta_p^2 = .05$ for negative example standards. There was no statistically significant interaction effect, $F(1, 122) = 1.68$, $p = .197$, $\eta_p^2 = .01$.

## Discussion

The present study investigated whether learners' judgment accuracy in evaluating self-generated examples can be enhanced by providing external standards during the example evaluation process. Specifically, we varied whether learners received expert example standards (with vs. without) and negative example standards (with vs. without) during the comparison with their own examples. In line with Hypothesis 1a and thus replicating the findings of Froese and Roelle (2022), we found that expert example standards significantly increased learners' judgment accuracy. Even though the main effects of expert examples were smaller than in Froese and Roelle's study, which in part could be due to the subadditive interaction effect that was observed, learners who were provided with expert example standards showed lower bias, better absolute and better relative accuracy than their counterparts. One explanation for these benefits of expert example standards is that expert examples helped learners recognize crucial features of quality examples for the respective concepts and hence form predictive cues they could use in judging their own examples. This, in turn, resulted in overall better judgment accuracy. Notably, like in Froese and Roelle (2022), the expert example standards did not yield fully accurate judgments. For instance, the learners still overestimated their poor examples, which were credited with 0% by the expert raters, by ca. 57%.

Contrary to our assumptions, the provision of negative example standards did not substantially contribute to (further) reducing the remaining inaccuracy in judging incorrect examples (Hypothesis 1b). Neither regarding learners' bias scores nor concerning the absolute accuracy of learners' judgments did we find a statistically significant main effect of negative example standards, and these standards even led to a substantial decrease in relative accuracy. The gamma correlation of $G = -.02$ basically indicates that the learners who received negative example standards were not at all able to discriminate between quality examples and not so well-designed examples. The fact that we did find a beneficial effect of negative example standards concerning bias when the group that only received negative example standards was compared to the group with no standards, indicates that parts of the lack of beneficial main effect can be attributed to a subadditive interaction effect, which occurred when expert example standards were provided as well (Hypothesis 1c).

Jointly, these findings could be explained as follows. The expert examples, which were beneficial, were designed such that the critical features of quality examples could be extracted by abstracting from the cover stories and hence focusing on the joint structural features (e.g., Gentner, 2010; Schalk et al., 2020). By contrast, although the negative example standards used different cover stories as well to prevent that learners focus on this surface feature, the respective negative examples scarcely shared common structural features. Rather, they were designed to cover typical errors of learner generated examples, which were identified in the data of Froese and Roelle (2022). Hence, the pairs of negative examples were not necessarily aligned in terms of structural features, which might have made it hard for learners to generate predictive cues for judging their own examples on this basis. Hence, unlike in the expert example standards, structural dissimilarity between the self-generated examples and the negative example standards was not easy to interpret. Specifically, structural dissimilarity between expert example standards and self-generated examples is a relatively valid cue for low quality, but structural dissimilarity between negative example standards and self-generated examples is not a valid cue for high quality, which made it hard for the learners to benefit from the negative example standards.

On this basis, one explanation for the subadditive interaction pattern could be that when learners received both types of standards, they mainly focused on the expert examples because, due to the above-mentioned reasons, these were easier to process and interpret than the negative example standards. When they did not also receive expert example standards, learners had to stick with the negative example standards and managed to slightly benefit from them, at least in terms of bias in comparison to the no standards group. This explanation keeps in line with our findings regarding subjective task difficulty (Hypothesis 2). We found that combining both types of standards did not increase the subjective difficulty of forming self-evaluations. However, both types of standards led to a moderate increase in learners' perceived mental effort, which seems to contradict our explanation that the learners mainly focused on the expert examples when they received both types of standards. However, the learners who received both types of standards did not experience higher mental effort in forming their judgments than the learners who received only one type of standard ($p = .438$ and $p = .371$, respectively), which could reflect that the learners mainly focused on the expert example standards even when they received both types of standards.

On a practical note, these results suggest preferring the provision of standards that represent a correct answer to a task assignment instead of incorrect or negative standards as only the former, at least for the task of evaluating self-generated examples, enable a meaningful comparison with the self-generated product through abstracting structural similarities. That is, despite the fact that negative examples did not exert the expected beneficial effects on learners' judgment accuracy, the present study enabled a replication of the *expert example standards effect* found by Froese and Roelle (2022), allowing to draw the conclusion that expert examples consistently function as an effective benchmark in evaluating self-generated examples. Naturally, it is nevertheless necessary to engage in further research that investigates the exact mechanisms behind the benefits of expert example standards (e.g., in think-aloud studies that delve into the comparison process more deeply) and replicates the effects with other learner populations and different learning material.

## Limitations and future research

The present study entails some important limitations that need to be considered. The first limitation refers to the focus of the study, which lies mainly on the effects of external standards on learners' judgment accuracy. Consequently, we neither implemented a measure that assessed learners' comprehension after they generated and evaluated the examples nor a subsequent regulation phase, in which learners could have adapted their examples. Investigating the effects of external standards on comprehension would have enabled to gain insight into potential explanatory approaches of how the standards exert their effects. For instance, an increased level of comprehension in the expert example groups would have hinted at a connection between accurate judgments and a deeper understanding of the learning material. A subsequent regulation phase, in which learners were given the opportunity to revise their examples or create new ones after the self-evaluation, would have provided insight into the degree to which the expert example standards can foster not only judgment accuracy but also effective regulation decisions in example generation. The absence of a learning outcome measure and regulation phase prevents conclusions concerning whether the provision of the expert example standards alone would be sufficient to help learners benefit from tasks in which they are engaged in example generation. Future studies should therefore extend the setting used in the present study and investigate not

only effects on judgment accuracy, but on regulation decisions and learning outcomes as well (see e.g., De Bruin et al., 2010; Thiede et al., 2003, 2017; Waldeyer & Roelle, 2023).

The second limitation relates to the design of the negative example standards. As outlined above, to support learners in finding similarities between their poor examples and the negative standards, we aligned the negative example standards with the poor examples that were created by the participants in Froese and Roelle's (2022) study, which entailed that the negative examples partly implied wrong illustrations, partly omitted components that were not illustrated at all, and partly also mixed-up components. Due to this design decision, however, it likely was hard for learners to extract critical structural features of poor examples from the standards and to interpret structural dissimilarity between their examples and the negative ones as a cue. It is thus an open question whether negative examples in general do not work or whether the inconsistent effects are due to the specific design used in the present study.

A third limitation refers to the role of example quality. In the present study as well as in previous research examining learners' judgment accuracy in evaluating self-generated examples (e.g., Froese & Roelle, 2022; Zamary et al., 2016), the quality of the learner generated examples was rather low. This low quality itself might affect judgment accuracy because it is well established that the solution rate of very difficult items is likely to be overestimated whereas the solution rate of very easy items is likely to be underestimated (e.g., Bromme et al., 2001). To investigate the role of example quality for the effects of both types of standards on judgment accuracy, we conducted exploratory moderation analyses that included example quality as a potential moderator for each of the eight declarative concepts for both absolute accuracy and bias (i.e., 16 moderation analyses in total). For the factor negative example standards, none of these analyses revealed a statistically significant interaction between example quality and the respective accuracy measure. For the factor expert example standards, in three of the 16 analyses we found a statistically significant interaction with example quality. When aggregated across the eight concepts, however, no statistically significant moderation effects were found. Jointly, these findings suggest that at least for the effects of expert example standards, example quality might be a potential moderator, though its effects likely are small and not consistent. Future studies that experimentally vary example quality (e.g., by the provision of expert examples beforehand) should test this tentative conclusion of the present study.

As a fourth limitation it is important to note that we did not assess how learners processed the external standards and how they used them in forming their judgments. Hence, it is not possible to confirm or falsify the above-mentioned assumptions that mainly the expert example standards were processed while the negative example standards were disregarded and that this could also explain the pattern of results concerning learners' perceived cognitive load in forming self-evaluations. It also remains unclear if the learners based their judgments on rather shallow cues (e.g., example length) or if they were able to abstract structural features of their examples that are predictive for their quality level. Methodological approaches such as self-explanations (e.g., Dinsmore & Parkinson, 2013) or think-aloud methods (e.g., Oudman et al., 2018) could capture the specific processes during the judgment process and provide indications to what extent the respective standards were processed by the learners.

Since we have not implemented measures that capture the specific processing of the external standards, we do not know whether the full potential of the standards was exploited or if there is still room for improvements. As we did not support the processing of the standards, the learners probably were not able to fully comprehend to what extent or why the respective idea units were or were not correctly illustrated in the presented standards. While

learners might have processed the standards on a rather shallow level due to missing support or prompts, deeper processing could be stimulated, for instance, by prompting the learners to color code the correct or incorrect idea units of the concepts in the standards in the first step, before color coding the respective components in their own examples in the second step. This procedure might foster learners' comprehension of the external standards and provide cues on the structural (dis-)similarity between the own example and the standards, which could enable the learners to generate more accurate judgments. Future studies could test this approach and examine if the enhancing effects of external standards on learners' judgment accuracy might intensify by inducing deeper processing of the standards.

## Conclusion

The present study implies three main contributions. First, the replication of the beneficial effects of expert example standards on learners' judgment accuracy concerning all three accuracy measures (i.e., bias, absolute and relative accuracy) strengthens the notion that expert example standards can function as an effective means to foster accurate self-evaluations in example generation tasks. Second, the lack of beneficial and partly (in terms of relative accuracy) even detrimental effects of negative example standards on learners' accuracy indicate that this type of negative standard should be avoided, which points to the third contribution. In line with the conclusions of Froese and Roelle's (2022) study, the present study indicates that external standards need to represent *correct* (and not incorrect) responses corresponding to a task assignment to effectively foster learners' judgment accuracy.

**Acknowledgements** We thank the research assistant Chantal Cebula for assisting in collecting and coding the data. Furthermore, we thank Dennie Froese for programming and piloting the experiment.

**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.

**Data availability** Not applicable.

**Code availability** Not applicable.

## Declarations

**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.

**Conflicts of interests/competing interests** The authors have no conflicts of interest to declare that are relevant to the content of this article.

## 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

Baker, J. M. C., Dunlosky, J., & Hertzog, C. (2010). How accurately can older adults evaluate the quality of their text recall? The effect of providing standards on judgment accuracy. *Applied Cognitive Psychology, 24*, 134–147. https://doi.org/10.1002/acp.1553

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

Bromme, R., Rambow, R., & Nückles, M. (2001). Expertise and estimating what other people know: The influence of professional experience and type of knowledge. *Journal of Experimental Psychology: Applied, 7*(4), 317–330. https://doi.org/10.1037/1076-898X.7.4.317

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

De Bruin, A. B. H., Camp, G., & Van Merriënboer, J. J. G. (2010). Available but irrelevant: when and why information from memory hinders diagnostic reasoning. *Medical Education, 44(*10), 948–950. https://doi.org/10.1111/j.1365-2923.2010.03789.x

De Bruin, A. B. H., Dunlosky, J., & Cavalcanti, R. B. (2017). Monitoring and regulation of learning in medical education: The need for predictive cues. *Medical Education, 51*, 575–584. https://doi.org/10.1111/medu.13267

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

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

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

Froese, L., & Roelle, J. (2022). Expert example standards but not idea unit standards help learners accurately evaluate the quality of self-generated examples. *Metacognition and Learning, 17*(2), 565–588. https://doi.org/10.1007/s11409-022-09293-z

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science, 34*(5), 752–775. https://doi.org/10.1111/j.1551-6709.2010.01114.x

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

Golke, S., Steininger, T., & Wittwer, J. (2022). What makes learners overestimate their text comprehension? The impact of learner characteristics on judgment bias. *Educational Psychology Review*. https://doi.org/10.1007/s10648-022-09687-0

Hiller, S., Rumann, S., Berthold, K., & Roelle, J. (2020). Example-based learning: Should learners receive closed-book or open-book self-explanation prompts? *Instructional Science, 48*(6), 623–649. https://doi.org/10.1007/s11251-020-09523-4

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Lachner, A., Hoogerheide, V., van Gog, T., & Renkl, A. (2022). Learning-by-teaching without audience presence or interaction: When and why does it work? *Educational Psychology Review, 34*(2), 575–607. 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

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110*(1), 40–48. https://doi.org/10.1037/0021-843X.110.1.40

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

Oudman, S., van de Pol, J., Bakker, A., Moerbeek, M., & van Gog, T. (2018). Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding. *Teaching and Teacher Education, 76*, 214–226. https://doi.org/10.1016/j.tate.2018.02.007

Paas, F. G. W. C. (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

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology, 88*(1), 144–161. https://doi.org/10.1037/0022-0663.88.1.144

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. (2015). Effects of comparing contrasting cases on learning from subsequent explanations. *Cognition and Instruction, 33*(3), 199–225. https://doi.org/10.1080/07370008.2015.1063636

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

Roelle, J., Schmidt, E. M., Buchau, A., & Berthold, K. (2017). Effects of informing learners about the dangers of making overconfident judgments of learning. *Journal of Educational Psychology, 109*(1), 99–117. https://doi.org/10.1037/edu0000132

Roelle, J., Froese, L., Krebs, R., Obergassel, N., & Waldeyer, J. (2022a). Sequence matters! Retrieval practice before generative learning is more effective than the reverse order. *Learning and Instruction, 80*, 101634. https://doi.org/10.1016/j.learninstruc.2022.101634

Roelle, J., Schweppe, J., Endres, T., Lachner, A., von Aufschnaiter, C., Renkl, A., Eitel, A., Leutner, D., Rummer, R., Scheiter, K., & Vorholzer, A. (2022b). Combining retrieval practice and generative learning in educational contexts: Promises and challenges. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie, 54*, 142–150. https://doi.org/10.1026/0049-8637/a000261

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

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

Schöne, C., Dickhäuser, O., Spinath, B., & Stiensmeier-Pelster, J. (2002). *Skalen zur Erfassung des schulischen Selbstkonzepts: SESSKO*. Hogrefe. https://doi.org/10.25656/01:2776

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. (2022). Domain-specific prior knowledge and learning: A meta-analysis. *Educational Psychologist, 57*(1), 31–54. https://doi.org/10.1080/00461520.2021.1939700

Steininger, T. M., Wittwer, J., & Voss, T. (2022). Classifying examples is more effective for learning relational categories than reading or generating examples. *Instructional Science, 50*(5), 771–788. https://doi.org/10.1007/s11251-022-09584-7

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66–73. https://doi.org/10.1037/0022-0663.95.1.66

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

Thiede, K. W., Redford, J. S., Wiley, J., & Griffin, T. D. (2017). How restudy decisions affect overall comprehension for seventh-grade students. *British Journal of Educational Psychology, 87*(4), 590–605. https://doi.org/10.1111/bjep.12166

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

Waldeyer, J., & Roelle, J. (2023). Does providing external standards after keyword generation improve metacomprehension accuracy and regulation for high school students? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*. https://doi.org/10.1026/0049-8637/a000266

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

Wylie, R., & Chi, M. T. H. (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. https://doi.org/10.1017/CBO9781139547369.021

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