Rubrics enhance accuracy and reduce cognitive load in self-assessment

Rebecca Krebs1,2 · Björn Rothstein3 · Julian Roelle1

Received: 18 May 2021 / Accepted: 20 April 2022 / Published online: 19 May 2022
© The Author(s) 2022

Abstract
Rubrics, which are designed to inform learners about assessment criteria and quality levels concerning a task assignment, are a widely used means to support learners in self-assessing their task performance. Usually, rubrics enhance task performance. Surprisingly, however, relatively little is known about the underlying mechanisms via which rubrics exert their beneficial effects. Although it is frequently suggested that the effects of rubrics on task performance are due to the fact that they support learners in accurately self-assessing their level of performance, which, in turn, paves the way for effective regulation, the empirical evidence that supports this notion is relatively scarce. Tightly controlled experimental studies, which allow isolating the effects of rubrics on the accuracy of self-assessments, are scarce. The present study was designed to close this gap. Specifically, in an experiment with \( N = 93 \) high school students we tested whether providing learners with a rubric in self-assessing the quality of self-written scientific abstracts would enhance judgment accuracy. As main results, we found that the rubric group showed higher absolute accuracy and lower bias. Furthermore, the rubric reduced cognitive load in self-assessing. These findings support the notion that one mechanism via which rubrics unfold their efficacy is by enhancing judgment accuracy.

Keywords Self-regulated learning · Rubrics · Metacognition · Self-assessment · Judgment accuracy · Cognitive load

Providing learners with rubrics, which are designed to inform them about assessment criteria and quality levels concerning a task assignment, is a widely used approach to implement self-assessment in educational practice (e.g., Panadero & Jonsson, 2020; Panadero et al., 2012). And that is a good thing – a large body of research indicates that rubrics enhance performance across educational levels (e.g., in primary school as well as in higher education) and various tasks such as writing texts or giving oral presentations (e.g.,

---

* Rebecca Krebs
rebecca.krebs@ruhr-uni-bochum.de

1 Faculty of Philosophy and Educational Research, Ruhr University, Bochum, Germany
2 Fakultät Für Philosophie Und Erziehungswissenschaft, Ruhr-Universität Bochum, Universitätsstraße 150, 44801 Bochum, Germany
3 Faculty of Philology, Ruhr University, Bochum, Germany
Andrade et al., 2009; Murillo-Zamorano & Montanero, 2018). Surprisingly, however, relatively little is known about the underlying mechanisms by which rubrics exert their beneficial effects. More specifically, although it is frequently suggested that one mechanism via which rubrics exert beneficial effects is that they support learners in accurately self-assessing their level of performance, which, in turn, enables learners to engage in effective self-regulation (e.g., Panadero & Romero, 2014), the empirical evidence that supports this notion is relatively scarce. Previous rubrics research was predominantly conducted in applied settings, in which it was not possible to isolate the effects of rubrics on the accuracy of forming self-assessments and hence to draw robust conclusions concerning this proximate ingredient of the benefits of rubrics.

The present study was designed to address this open issue. In a tightly controlled experiment, we manipulated whether high school students, who had written a scientific abstract about a previously read scientific article, were supported by a rubric in assessing the quality of their abstracts. Absolute accuracy, bias, and relative accuracy of the self-assessments as well as subjective cognitive load during the formation of self-assessments were used as the main dependent variables.

### Rubrics: How do they exert beneficial effects?

Rubrics are documents (often: tables) that are designed for the specific purpose of informing learners about the expectations of a task assignment. Specifically, rubrics provide learners with the assessment criteria and, based on these criteria, illustrate different levels of quality regarding performance on the respective task assignment (e.g., Panadero & Jonsson, 2013; Reddy & Andrade, 2010). In view of their obvious scaffolding character with respect to the formation of assessments, rubrics are widely used in the assessment of (complex) performances (e.g., Andrade Goodrich & Boulay, 2003; Jonsson, 2014). Furthermore, in recent years the implementation of rubrics as a tool for self-assessment has received increasing attention. Across various educational levels (e.g., in primary school as well as in higher education, see Lerdpornkulrat et al., 2019; Smit et al., 2017) and substantially different task assignments such as presenting academic content (e.g., Murillo-Zamorano & Montanero, 2018), learning about statistics (e.g., Jonsson, 2014), or writing (scientific) texts (e.g., Andrade Goodrich & Boulay, 2003; Ashton & Davies, 2015; Nordrum et al., 2013), rubrics are widely used as a means for supporting self-assessment on part of the learners (e.g., Panadero & Romero, 2014; see also Panadero et al., 2017).

The widespread use of rubrics occurs against the backdrop of robust empirical evidence concerning its benefits. Although under certain circumstances (e.g., strict time constraints) rubrics might induce stress and pressure and thus hinder task performance (Panadero & Romero, 2014), overall meta-analyses and narrative reviews clearly indicate beneficial effects of rubrics on the application of self-regulated learning strategies, self-efficacy, motivation and task performance (e.g., Panadero & Jonsson, 2013; Panadero et al., 2017). In terms of the underlying mechanisms that lead to these beneficial effects, however, the evidence is less robust and conclusive. From a theoretical perspective, one proximate active ingredient of the benefits of rubrics is that they clarify the assessment criteria and components of quality performances on the respective task assignments (e.g., Goodrich Andrade, 2001; Schamber & Mahoney, 2006; for an overview see Panadero & Jonsson, 2013). This clarification of assessment criteria and quality levels is expected to increase the degree to which learners base their self-assessments on cues of high diagnosticity (i.e.,...
the criteria that are communicated in the rubric). In line with cue utilization theory (e.g., Koriat, 1997; see also De Bruin & Van Merrienboer, 2017), this increase in reliance on diagnostic cues, in turn, should increase the accuracy of learners’ self-assessments, which, in line with metacognitive models of self-regulated learning (e.g., Nelson & Narens, 1994; see also De Bruin et al., 2020), finally results in effective regulation.

From an empirical view, however, the evidence concerning this potential mechanism via which rubrics exert their beneficial effects is scarce for several reasons. First, most previous rubrics studies focused on rubrics that were used in a formative manner, which means that the learners could use the rubric not only during self-assessing their performance, but also during the execution of the respective task (e.g., Panadero & Jonsson, 2013). For instance, Lipnevich et al. (2014) gave students the opportunity to revise an assignment that was part of a course by using a rubric (i.e., the learners could use the rubric in self-assessing and in revising their drafts). Similarly, Andrade et al. (2010) involved the students in the process of generating criteria for an effectively written essay and reviewing a written rubric before using the rubric to revise their own work. Consequently, in these studies the rubrics could affect not only the accuracy of learners’ self-assessments but also task performance (e.g., due to effective regulation, see Bradford et al., 2016; Panadero & Romero, 2014). Although such potential indirect effects of rubrics on task performance are desirable from an applied point of view, they are obstructive in investigating the effects of rubrics on self-assessment accuracy and thus the potential underlying mechanism via which rubrics exert their benefits. Given that learners frequently overconfidently judge their level of performance and learning (e.g., Dunlosky & Rawson, 2012; Dunning et al., 2003; Froese & Roelle, 2022; see also Prinz et al., 2020b), increases in task performance per se can result in increased self-assessment accuracy. Hence, when rubrics enhance task performance, it is hard to draw robust conclusions concerning whether rubrics actually directly increased the accuracy of the self-assessment, or whether effects on self-assessment accuracy are (simply) due to the beneficial effects on task performance.

A second issue in previous rubrics research that hinders robust conclusions concerning the effects of rubrics on self-assessment accuracy is that with very few exceptions the studies did not implement tightly controlled experimental designs (see Panadero & Jonsson, 2013; for a rare exception, see Panadero et al., 2012). Rather, quasi-experimental and descriptive studies dominate the recent rubrics literature (Lerdpornkulrat et al., 2019; Panadero & Romero, 2014; Smit et al., 2017), which is understandable given the strong emphasis on application and relevance for classroom instruction in this field of research, but detrimental in terms of being able to clearly attribute the reported effects to the provision of rubrics. A third problematic issue that is closely related to the second one is that previous rubrics studies that implemented quasi-experimental designs sometimes manipulated more than one aspect between conditions and hence whole treatment packages rather than single components were evaluated. For instance, in Murillo-Zamorano and Montanero (2018), students not only worked with a rubric, but also gave or received feedback from a peer as part of a peer assessment with rubric activity. Although the investigation of treatment packages is desirable from the perspective of applied research, the effects that are reported in the studies that follow this procedure are hard to attribute to the provision of rubrics alone.

In view of these circumstances in the rubrics literature, it would be worthwhile to investigate whether rubrics actually directly affect the accuracy of learners’ self-assessments. At first glance, however, one could argue that this issue is already well investigated, even if not under the label rubrics. Specifically, one could argue that the literature on standards essentially already clearly shows that rubrics exert beneficial effects on self-assessment
accuracy. Basically, standards are external representations of the correct answer to a task assignment (e.g., Baars et al., 2014; Dunlosky & Lipko, 2007; Rawson & Dunlosky, 2007). For instance, when a task assignment requires learners to retrieve concept definitions from memory, providing learners with the correct definitions of the concepts would be conceived as a standard (see Baker et al., 2010; Rawson & Dunlosky, 2007). Likewise, the correct solutions to problem-solving steps are conceived as standards (e.g., Baars et al., 2014). Such 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; see also Baker et al., 2010; Dunlosky et al., 2011; Nederhand et al., 2018) and extended to problem-solving tasks by Baars et al. (2014). Lipko and colleagues also further developed the format of the standards. Specifically, Lipko et al. developed so-called idea unit standards by parsing the correct answer to the respective task into its constituent idea units and, potentially because it highlights the crucial components of the target response more clearly, could show that this format has added value to providing learners with full standards that are not divided into idea units.

The outlined idea unit standards in particular could be conceived as being highly similar to rubrics at first glance, because, like rubrics, they provide learners with an overview of the assessment criteria to a task assignment and hence can serve as scaffolding for the formation of self-assessments. However, at second glance there are also substantial differences between (idea unit) standards and rubrics that need to be considered. First, what is compared by the learners differs substantially between idea unit standards and rubrics. Idea unit standards explicitly include the correct answer to a task and thus allow relatively straightforward direct comparisons of learners’ products and the standards. For instance, learners can directly check whether a definition they retrieved from memory contains the provided idea units of the correct definition. By contrast, rubrics do not include the correct answer to a task, but rather inform learners about the rather abstract criteria that a correct answer needs to meet. Hence, learners cannot directly compare their products to the criteria provided by the rubric, but need to infer whether their products meet the criteria. For instance, in the task of writing a scientific abstract to a text, a rubric would not include the correct abstract to the text, but rather inform learners about the criteria that a quality abstract would need to meet (e.g., in terms of structure or scientific style). A second important difference between idea unit standards and rubrics, which is related to the outlined differences concerning what is compared, relates to the granularity of the comparisons. In the literature on standards, learners are usually merely required to judge whether an idea unit is or is not covered by the generated products (e.g., Lipko et al., 2009; Zamary et al., 2016). By contrast, rubrics often require learners to make comparisons that are more fine-grained such as deciding which of four quality levels concerning a specific criterion they have reached (e.g., Panadero & Romero, 2014, see also Hafner & Hafner, 2003). Furthermore, rubrics sometimes even require learners to compare their ratings to various criteria in order to find out their strengths and weaknesses concerning a task assignment (e.g., Andrade Goodrich & Boulay, 2003).

In view of these substantial differences, it would be quite a leap to generalize the evidence concerning the benefits of idea unit standards regarding judgment accuracy to rubrics. Although both support measures require learners to engage in comparing their products with external benchmarks, the inferential level and the granularity of the required comparisons is considerably higher for rubrics, which probably also reflects the fact that the complexity of the task assignments for which the respective measures are used is usually higher for rubrics than
for (idea unit) standards (e.g., writing of scientific texts vs. retrieval of concept definitions). Nevertheless, the literature on standards can inform research on rubrics, for it points to an aspect that has widely been ignored in the rubrics literature. Specifically, the standards literature highlights that effective means to foster self-assessment accuracy need to deal with learners’ limited working memory capacity and hence manage cognitive load during the formation of the self-assessments (e.g., Dunlosky et al., 2011).

In self-regulated learning settings, learners need to devote cognitive capacity not only to performing the respective task at hand, but also to accurately monitoring their task performance (e.g., De Bruin et al., 2020). As cognitive capacity is limited (e.g., Sweller et al., 2019) and as learners who perform a complex task for the first time in particular usually experience substantial cognitive load while performing the task, it often happens that little to no working memory capacity is left for the execution of monitoring processes. This, in turn, hinders the formation of accurate self-assessments (e.g., Panadero et al., 2016; see also Kostons et al., 2009, 2012). On this basis, it is reasonable to assume that the benefits of instructional means to enhance self-assessment accuracy in self-regulated learning (such as rubrics) depend on the demands they place on working memory resources. More specifically, these means should be promising when they reduce rather than increase the cognitive load in forming self-assessments in particular. It is an open question, however, how rubrics would affect cognitive load during self-assessment and whether their effects on self-assessment accuracy would be meaningfully related to the potential effects on cognitive load. On the one hand, it could be argued that rubrics might contribute to increases in cognitive load because they require learners to infer to what degree their products meet the criteria and quality levels described in the respective rubric. On the other hand, they might decrease cognitive load because in comparison to having no guidance at all, they prevent learners from (randomly) generating and trying cues that could be used in assessing their task performance.

Importantly, investigating the effects of rubrics on cognitive load is interesting and relevant not only from a basic research perspective. Rather, investigating effects of rubrics on cognitive load could also be informative in terms of evaluating the potential of rubrics in authentic educational settings. Instructional means that put excessive cognitive demands on learners likely would not be useful in authentic settings even if they yield beneficial results in tightly controlled laboratory settings, because learners likely would avoid them when they can decide on their own in which activities they engage. For instance, likely because they induce high cognitive load (and produce undesirable effects regarding motivation) in self-regulated journal writing learners tend to avoid engaging in sophisticated metacognitive learning strategies (Nückles et al., 2020, see also Moning & Roelle, 2021). Given that rubrics are designed to scaffold the formation of self-assessments, as outlined above they could be expected to reduce rather than increase cognitive load in self-assessment and hence the outlined detrimental effects potentially would not apply to the provision of rubrics. Nevertheless, as there are also reasons to believe that rubrics could increase cognitive load in self-assessment, experimental evidence concerning the effects of rubrics on cognitive load is needed to evaluate these potential undesired side-effects of rubrics.

The present study

Against the outlined theoretical and empirical background, the main goal of the present study was to analyze the effects of rubrics on self-assessment accuracy in a tightly controlled experimental setting. Specifically, we addressed the following research questions.
Research Question 1: Do rubrics enhance self-assessment accuracy (i.e., enhance the product of self-assessment)? The research reviewed above suggests that rubrics exert beneficial effects on self-assessment accuracy, but rigor tests of these effects are scarce. Hence, we tested the following directional hypothesis:

Hypothesis 1. Students with a rubric show higher self-assessment accuracy than students without rubric.

Research Question 2: Do rubrics affect subjective cognitive load in forming self-assessments (i.e., one potential important ingredient of beneficial effects of rubrics on judgment accuracy)? As outlined above, there are reasons to assume that rubrics decrease cognitive load as well as reasons to believe that rubrics increase cognitive load in self-assessment. Hence, we tested the following non-directional hypothesis:

Hypothesis 2. Learners with a rubric differ from learners who do not receive a rubric in cognitive load during self-assessment.

Research Question 3: Does the potential beneficial effect of rubrics on self-assessment accuracy depend on the inferential level of the assessment criteria? As a largely explorative secondary goal, we were furthermore interested in whether the outlined potential effects of rubrics on the products of learners’ self-assessment (i.e., increased accuracy) would depend on the inferential level of the assessment criteria. An essential component of rubrics is the detailed description of the assessment criteria and of several levels of quality for a task assignment (Andrade & Du, 2005). Albeit, unlike idea unit standards (see above), none of the criteria that are included in rubrics represent the correct answer to the respective tasks and thus allow direct comparisons with learners’ products, the degree to which some criteria are met can be determined via relatively parsimonious inferences (e.g., whether a scientific abstract includes information on the participants of a study or whether information on the participants is reported before the findings are described), whereas other criteria require more substantial inferences on part of the learners in order to determine whether the criteria are met (e.g., whether a scientific abstract is written in scientific style). More specifically, in self-assessing performance on complex tasks (e.g., writing tasks) in particular, such high inferential criteria likely are the rule rather than the exception. For instance, Andrade et al. (2009) provided learners with a rubric for writing that consisted of three global criteria (ideas and content, organization, and paragraphs) as well as four finer grained criteria (voice and tone, word choice, sentence fluency, and conventions). At least the four fine-grained criteria (but probably also the three global criteria) likely required learners to generate substantial inferences in self-assessing whether their products met the criteria (e.g., it is not possible to simply observe the degree to which voice and tone or sentence fluency was adequate). Similarly, these criteria have been implemented in other studies about writing as well (e.g., see Andrade, et al., 2010; Sadler & Good, 2006). But also in other domains, for example in nursing students’ clinical reasoning during simulated patient care scenarios (Jensen, 2013) or in teaching clinical reasoning in physical therapy training (Furze et al., 2015), high inferential criteria (e.g., whether learners can critically self-reflect on their reasoning ability or are able to incorporate complex factors in addressing patient needs) are common.

It is unclear how much learners can benefit from rubrics that involve high inferential criteria, at least when they do not receive excessive training on how to use the respective rubric. In the study by Andrade et al., which indeed showed significant beneficial effects,
the learners were familiarized with the rubrics in a relatively time-consuming manner. Given time constraints in education, however, such extensive familiarization likely happens quite rarely. In these conditions, for low inferential assessment criteria rubrics likely still work well, because they require relatively little introduction. For high inferential criteria, by contrast, rubrics might be less efficient. More specifically, it is unclear whether a relatively brief introduction into the respective criteria would be sufficient to realize substantial effects on self-assessment accuracy. Rather, for high inferential criteria learners might need substantial amounts of training (similar to interrater agreement training in scientific studies) before they can use them in forming accurate self-assessments. Hence, we tested:

Hypothesis 3. The potential beneficial effect of rubrics on self-assessment accuracy decreases as the required inferential level increases. That is, the benefits of rubrics should be more pronounced for low inferential than for high inferential criteria.

Method

Sample and design

The participants of this experiment were \( N=93 \) eleventh- and twelfth-grade students from two German high schools (49 female, 44 male). They were between 15 and 18 years old (\( M=16.74, \ SD=0.78 \)). Written informed parental consent was given for all students under age 18 and own informed consent for all students at the age of 18 or older.

The study followed an experimental between-subject design with two conditions (random assignment). In an initial study phase, all participants received an introduction in writing scientific abstracts as well as in four criteria for judging the quality of scientific abstracts. Afterwards, all students read an adapted (i.e., shortened) version of a real scientific article, which did not include an abstract, and were then asked to write an abstract for the article. After the writing phase, the experimental manipulation was carried out. The participants in the rubric condition (\( n=46 \)) received a rubric that provided support in rating the level to which each of the four previously introduced criteria were met in their abstracts. The participants in the control condition (\( n=47 \)), who were also introduced to the four criteria before they wrote their abstracts, rated the level to which the four criteria were met without being supported by a rubric.

Materials

Except for the scientific article (see below), all materials were provided in a computer-based learning environment. All materials were designed in cooperation with an expert in the field of German didactics (i.e., the second author of this manuscript).

Introduction in writing a scientific abstract and in criteria for quality abstracts

All participants received an introduction in how to write a scientific abstract. In this introduction, they were first familiarized with what an abstract actually is and which functions it fulfills. The students got to know that an abstract is a short, factual summary of a scientific article, which sums up the most important aspects of that article. According to Swales (1990), Swales and Feak (2009) and Busch-Lauer (2012) the following parts of an abstract...
were mentioned: Introduction with question or hypotheses, method, results and discussion. As the function of an abstract, the students learned that it is the first contact of the reader with the article and should help to decide whether they are interested in reading the whole article.

In the next step the students were introduced in how to write an abstract. The explanations were structured according to four different criteria: (1) content, (2) sentence structure and grammar, (3) scientific style and (4) structure of the abstract. The criterion of content explained in detail which aspects should be mentioned in an abstract. Specifically, the students were told that an abstract should include a brief description of the problem or question the article is about, the subjects and/or data basis with any important characteristics, the method, the key findings of the experiment and a main conclusion that can be drawn from the experiment. The second criterion, sentence structure and grammar, focused on the linguistic style of a scientific abstract. Here, the students were informed that an abstract is largely structured and written like a report (which is a basic text type the students already knew how to write) and thus should be written in a concise, clear manner. The next criterion the students got introduced to was the scientific style of an abstract. Here, the students learned that an abstract’s language is precise and as objective as possible. Finally, the criterion of structure was mentioned. Here, the students were informed about the structure an abstract is mostly written in. This criterion was very close to the criterion of content: the students were again made aware of what parts (problem or question, data basis, methods, results and conclusion) should be present in an abstract and in what order.

The descriptions of criteria show that in contrast to the criteria of content and structure, the criterion of sentence structure and grammar as well as the criterion of scientific style cannot easily be inferred. Learners can relatively easy check whether they have mentioned the method, or whether they have written the method section before the results and after the description of the data basis. But as described above, in writing tasks rubrics often involve high inferential criteria such as the criteria of sentence structure and scientific style, which require a certain amount of inferencing on part of the learner.

To support learners’ understanding of the criteria, the students received two solved example problems that included high quality abstract passages for each aspect. An overview about the aspects and examples for the criterion of content can be found in Fig. 1. According to research in the field of example-based learning, providing multiple (solved) example problems usually substantially fosters initial cognitive skill acquisition (e.g., Hübner et al., 2010; Roelle et al., 2012; Schworm & Renkl, 2007; for an overview, see Renkl, 2014). In addition to the provision of solved example problems that showed high-quality passages, the learners were also provided with some examples of poor-quality passages. More specifically, in view of the notion that contrasted examples can help learners discern the critical features in solved example problems (e.g., Gentner, 2010; Quilici & Mayer, 1996; Schalk et al., 2020), we contrasted high-quality with poor-quality examples to highlight the critical features of quality abstracts. Figure 2 shows the contrasting cases for the criterion of scientific style.

**Scientific article**

To create a realistic basis for writing a scientific abstract, after they were introduced to the writing of abstracts the students were provided with a slightly adapted version of a real scientific article before they wrote their abstracts. Specifically, the article we used was published under the title “Steigert Kaugummikauen das kognitive Leistungsvermögen? Zwei Experimente der
"Does Chewing Gum Enhance Cognitive Abilities? Two Rather Extraordinary Experiments" by Rost et al. (2010) and dealt with the potential benefits of chewing gum regarding different measures of cognitive performance. The participants read the article, as it was published, in German. Since the participants had little to no statistical prior knowledge, we adapted the article such that the results could be understood without any knowledge about statistics. Furthermore, because two experiments were described in the original article, which increases the difficulty of writing a coherent abstract, the article was restructured so that the paper could be read as one experiment. The shortened article was comprised of 777 words.
and followed the established structure of scientific articles (Introduction, Method, Results, Discussion).

Rubric and self-assessment

The rubric was related to the four criteria of quality abstracts that all students had learned about in the beginning. For each criterion, the rubric provided specific guidance regarding the aspects that need to be met for all six levels (1: very good, 6: very poor) of the Likert-scales on which the learners were required to judge the quality of their abstracts. The aspects that needed to be met for each criterion were determined in cooperation with two university professors with ample experience in reviewing scientific articles.

During self-assessing their abstracts, the learners who received rubrics could review their abstracts, which were displayed in a text box on the right side of the screen, whenever and as long as they wanted. They received the prompt to “Please rate your abstract according to criterion content/ sentence structure and grammar/ scientific style/ structure and enter your rating in the text box.” and took their self-assessments by clicking on one of the six quality levels of the respective Likert-scale that was provided in conjunction with the rubric (see Fig. 3). After taking a self-assessment and before navigating to the next criterion, the learners were informed that they could not go back to the previous rating.

The learners in the control condition received the same rating scales for each criterion, the same prompt and could also review their abstracts while they judged the quality of their abstracts. The only difference was that the control condition learners did not receive any guidance concerning the aspects that had to be fulfilled for each of the six levels of the scales. A screenshot of the rating procedure in the control condition is provided in Fig. 4.

![Fig. 3 Self-assessment for the criterion of content (rubric condition)](image_url)
In order to determine the accuracy of students’ self-assessments regarding the four criteria, two experts (i.e., one psychology professor who serves on the editorial board of several scientific journals and one research assistant) used the rubric that was also given to the students in the rubric condition and provided ratings concerning the four above-mentioned criteria. The two experts independently judged 20 abstracts. The interrater reliability, as determined by the intraclass correlation coefficient with measures of absolute agreement, was very good (all ICCs > 0.85). In view of the very good interrater reliability, the remaining abstracts were judged by one rater only.

Self-assessment accuracy

In order to determine self-assessment accuracy, we used the students’ assessments regarding the four criteria and the expert ratings. Specifically, we used these ratings to determine three different measures of assessment accuracy that were described by Schraw (2009). First, we calculated absolute accuracy, which was operationalized as the difference between the students’ assessments and the experts’ ratings regardless of the direction of the difference (i.e., |X_{Student’s Judgment} – X_{Expert’s Judgment}|). Hence, positive and negative differences are both counted as inaccuracies and do not cancel each other out.

Second, we determined bias, which was operationalized as the signed difference between students’ assessments regarding the four criteria and the respective expert ratings (i.e., X_{Student’s Judgment} – X_{Expert’s Judgment}). The bias measure allows for measuring students’ over- and underconfidence with positive values indicating that students overestimated their performance, and negative values indicating that students underestimated their performance.
Third, we determined relative accuracy, which refers to the degree to which the students can discriminate between their performance concerning the four criteria. For this purpose, we calculated intra-individual correlations between the students’ performance and their self-assessments. Specifically, we computed Goodman–Kruskal’s gamma correlations (Goodman & Kruskal, 1954), whereby stronger positive correlations indicate greater relative accuracy. Gamma correlations are the established measure in metacomprehension research for determining relative accuracy (e.g., De Bruin et al., 2011; Thiede et al., 2003; Waldeyer & Roelle, 2021). Note that if there is no variability in learners’ judgments or performance concerning the four criteria, the gamma correlation cannot be determined. Learners with such tied scores (11 learners in the conditions with rubrics and 12 learners in the condition without rubrics) were hence not included in the analyses concerning relative accuracy.

Cognitive load

After self-assessing their abstracts, the students were asked to rate the mental effort they invested in their self-assessment and the perceived difficulty of forming their self-assessments on 7-point Likert scales (1: very very low mental effort/difficulty, 7: very very high mental effort/difficulty). The wording of the questions 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).

Procedure

Each student worked individually (most of the time: on a personal computer) throughout the entire experiment. First, all students gave general information on their gender and age as well as their last grade in the subject German. Second, the students received the thematic introduction in how to write abstracts. Afterwards, they were asked to read the scientific text, which was available in paper form on each table. Reading time was held constant at 20 min. Next, they were asked to write an abstract for the article. At the end of the writing time, which was held constant at 20 min for all participants, the students had to hand the article back to the researcher and then were asked to rate their abstracts concerning the four criteria. During forming the self-assessments, all learners could access their self-written abstracts (but not the article itself), which were displayed next to the respective rating scale. The students had as much time as they needed for their self-assessments. Finally, the students answered the mental effort and perceived difficulty items. The experiment lasted approximately 90 min.

Results

The mean scores and standard deviations for both groups on all measures are shown in Table 1. We used an α-level of 0.05 for all tests. No participants were excluded from the analyses. For effect sizes, we report Cohens’ $d$ for $t$ tests and partial $\eta^2$ for $F$ tests. Based on Cohen (1988), values around $d=0.20$ and $\eta_p^2=0.01$ can be considered as small effects, values around $d=0.50$ and $\eta_p^2=0.06$ as medium effects, and values around $d=0.80$ or $\eta_p^2=0.14$ or higher as large effects.

 Springer
Preliminary analyses

In the first step, we tested whether the random assignment resulted in comparable groups. We did not find a statistically significant difference between the groups regarding age, $t(90) = -0.03$, $p = 0.792$, $d = -0.05$, grades in the subject German, $t(90) = 0.54$, $p = 0.593$, $d = 0.11$ or gender, $\chi^2(1) = 1.03$, $p = 0.597$. As the abstracts were written before the experimental manipulation took place, we also tested whether the abstracts were comparable in length and quality. Neither regarding the number of written words, $t(91) = 0.15$, $p = 0.878$, $d = 0.03$, nor regarding any of the four criteria against which the quality of the abstracts were judged by the expert raters, did we find any statistically significant differences between the groups, $t(91) = 0.63$, $p = 0.529$, $d = -0.13$ for content, $t(91) = 0.99$, $p = 0.325$, $d = 0.20$ for sentence structure, $t(91) = 0.46$, $p = 0.648$, $d = -0.09$ for scientific style, $t(91) = -0.53$, $p = 0.596$, $d = 0.11$ for structure, respectively. Jointly, these findings indicate there were no significant a priori differences in learning prerequisites between the experimental groups. Furthermore, the findings concerning abstract quality indicate that the effects of the rubric on judgment accuracy (see below) cannot be attributed to effects of the rubric on task performance.

Table 1  Means and standard deviations of all measures

| Measure                                      | Rubric condition | Control condition |
|----------------------------------------------|------------------|-------------------|
| Age (in years)                               | 16.72 (0.83)     | 16.76 (0.74)      |
| German grade*                                | 2.85 (0.99)      | 2.74 (0.95)       |
| Word count                                   | 151.93 (38.10)   | 150.74 (36.31)    |
| Expert judgment: content                     | 1.30 (0.63)      | 1.23 (0.43)       |
| Expert judgment: sentence structure          | 1.98 (0.88)      | 1.79 (0.98)       |
| Expert judgment: scientific style             | 2.02 (1.02)      | 1.94 (0.76)       |
| Expert judgment: structure                   | 1.13 (0.34)      | 1.17 (0.38)       |
| Self-assessment: content                      | 5.21 (0.99)      | 4.59 (1.05)       |
| Self-assessment: sentence structure           | 4.71 (0.91)      | 4.55 (1.06)       |
| Self-assessment: scientific style             | 4.74 (0.88)      | 4.59 (0.65)       |
| Self-assessment: structure                   | 4.74 (0.83)      | 4.91 (0.93)       |
| Absolute accuracy                             | 0.84 (0.52)      | 1.05 (0.60)       |
| Absolute accuracy high inferential            | 0.97 (0.53)      | 1.05 (0.72)       |
| Absolute accuracy low inferential             | 0.71 (0.70)      | 1.05 (0.80)       |
| Relative accuracy                             | 0.58 (0.68)      | 0.41 (0.76)       |
| Bias                                         | -0.40 (0.75)     | -0.81 (0.80)      |
| Bias high inferential                         | -0.27 (0.87)     | -0.56 (1.11)      |
| Bias low inferential                          | -0.53 (0.83)     | -1.03 (0.83)      |
| Mental effort                                 | 3.33 (1.49)      | 3.57 (1.49)       |
| Perceived difficulty                          | 2.63 (1.16)      | 3.28 (1.38)       |

*In German high schools 1 is the best and 6 the worst grade that can be attained
**Self-assessment accuracy**

To test our hypothesis that rubrics would increase self-assessment accuracy (Hypothesis 1), we compared the groups concerning absolute accuracy, bias and relative accuracy. In terms of absolute accuracy, a *t* test indicated a statistically significant difference, *t*(90) = -1.85, *p* = 0.034 (one-sided), *d* = 0.39. The rubric group showed better absolute accuracy than the control group, that is, values of the rubric group deviated less from the expert values than those of the control group. With respect to bias, we found a similar pattern of results. Both groups showed underconfidence, but the rubric group’s assessments were less biased than the control group’s assessments, that is, the rubric group’s bias scores were closer to zero, *t*(90) = 2.46, *p* = 0.008 (one-sided), *d* = 0.52. In terms of relative accuracy, by contrast, we did not find a statistically significant difference between the groups, *t*(67) = -0.85, *p* = 0.198 (one-sided), *d* = 0.21.1 The relatively high gamma correlations (0.58 in the rubric group and 0.41 in the control group) indicate that both groups were fairly able to discriminate their performance concerning the four criteria.

Beyond the overall effect of the rubric on self-assessment accuracy, we were also interested in whether the potential beneficial effect of rubrics on self-assessment accuracy would depend on the inferential level required by the criteria. Specifically, in Hypothesis 3 we assumed that the potential beneficial effect of rubrics on self-assessment accuracy would decrease as the required inferential level increases. To test this hypothesis, we conducted mixed-repeated measures ANOVAs with condition (rubric vs. no rubric) as between-subjects factor and inferential level of the criteria (low inferential level vs. high inferential level) as within-subjects factor.2 In terms of absolute accuracy, the ANOVA did not reveal a statistically significant effect of inferential level, *F*(1, 90) = 2.41, *p* = 0.124, *η*^2^ = 0.02. More importantly, there was also no statistically significant interaction between condition and inferential level, *F*(1, 90) = 2.41, *p* = 0.124, *η*^2^ = 0.02, which indicates that the effects on absolute accuracy did not (substantially) depend on the inferential level of the criteria. We found a similar pattern of results concerning bias. Although the ANOVA indicated a statistically significant effect for inferential level, *F*(1, 90) = 12.71, *p* = 0.001, *η*^2^ = 0.12, which reflects that, surprisingly, bias was substantially lower for the high inferential level criteria than for the low inferential level criteria, there was no statistically significant interaction between condition and inferential level, *F*(1, 90) = 0.94, *p* = 0.333, *η*^2^ = 0.01. Hence, the beneficial effects of the rubric regarding bias did not depend on the criteria’s inferential level.

**Exploratory analyses**

Despite the above-mentioned substantial differences, rubrics share some commonalities with idea unit standards as both support measures serve as a type of external benchmark. Hence, in view of the notion that the benefits of idea unit standards are most pronounced for learners whose task performance is low (see Nederhand et al., 2018), for exploratory purposes we analyzed whether the effects of rubrics on absolute accuracy, bias and relative

---

1 Please note that the gamma correlations for *n* = 23 participants could not be determined due to invariant judgments or performances on the four criteria.

2 We did not use relative accuracy as dependent variable in these analyses, because as we had only four criteria, it was hardly possible to determine separate meaningful gamma correlations for the two low inferential level and the two high inferential level criteria.
accuracy would depend on learners’ task performance. For none of the accuracy measures, however, the interaction between condition and learners’ overall performance in writing the abstracts was statistically significant, $F(1, 88) = 1.50, p = 0.224, \eta_p^2 = 0.01$ for absolute accuracy, $F(1, 89) = 1.23, p = 0.269, \eta_p^2 = 0.01$ for bias, and $F(1, 66) = 1.81, p = 0.183, \eta_p^2 = 0.02$ for relative accuracy.

**Cognitive load**

In Hypothesis 2, we assumed that learners with rubric would differ from learners without rubric concerning cognitive load during self-assessment. To test this hypothesis, we compared the groups concerning perceived difficulty and mental effort. In terms of perceived difficulty, a $t$ test indicated a statistically significant effect, $t(90) = -2.45, p = 0.016, d = 0.52$. The rubric group perceived forming the self-assessments as easier than the control group. By contrast, in terms of mental effort, we did not find a statistically significant difference between the two groups, $t(90) = -0.77, p = 0.444, d = 0.16$.

**Exploratory analyses**

For exploratory purposes, we conducted mediation analyses to test whether the reduction in terms of perceived difficulty drove the beneficial effects of rubrics on absolute accuracy and bias. A mediation analysis tests if the effect of variable X (here: rubric vs. no rubric) on an outcome Y (here: measures of judgment accuracy) can be explained by the effect of X on the mediator M (here: perceived difficulty), which then affects Y. We conducted the mediation analyses with 95% percentile bootstrap confidence intervals from 10,000 bootstrap samples using the SPSS macro PROCESS (see Hayes, 2013). The results of the mediation analyses are shown in Figs. 5 and 6. Neither in terms of the effect on absolute accuracy, nor in terms of the effect on bias, did we find a statistically significant indirect effect via perceived difficulty (i.e., zero was included in the confidence intervals), $a_1 \times b_1 < 0.001 [-0.068, 0.066]$, and $a_2 \times b_2 = -0.030 [-0.125, 0.052]$. These findings indicate that the effect on perceived difficulty did not mediate the benefits of rubrics concerning absolute accuracy and bias.

We also explored whether the effects of rubrics on the two measures of subjective cognitive load would depend on learners’ task performance. For none of the measures, however, the interaction between condition and learners’ overall performance in writing the...
abstracts was statistically significant, $F(1, 88) = 1.83, p = 0.178, \eta_p^2 = 0.02$ for subjective difficulty and, $F(1, 88) < 0.01, p = 0.953, \eta_p^2 < 0.01$ for mental effort.

**Discussion**

The rubric beneficially affected learners’ self-assessments. First and foremost, in terms of absolute accuracy and bias the rubric fostered self-assessment accuracy (Hypothesis 1). One theoretical explanation for these benefits is that by transparently communicating the quality levels, the rubric provided the learners with diagnostic cues on which they could base their self-assessments. Hence, in contrast to the learners without the rubrics, the learners with rubrics were provided with cues that were highly predictive of their actual performance, which, in line with cue-utilization theory (e.g., Koriat, 1997; see also De Bruin & Van Merriënboer, 2017) fostered judgment accuracy. The finding that the rubrics furthermore decreased subjective cognitive load supports this conclusion (Hypothesis 2). Specifically, the decreased subjective difficulty of forming self-assessments that was reported by the learners who received rubrics indicates that rubrics serve an offloading function in that they relieve learners from deciding on their own which assessment criteria would need to be met for the different quality levels. However, the lower perceived difficulty did not serve as a mediator of the effects on judgment accuracy. Although at first glance this finding contradicts the notion that the lowering of subjective task difficulty is an important ingredient of beneficial effects of rubrics on judgment accuracy, at second glance this finding likely indicates that the reduction of cognitive load is probably rather “only” an important prerequisite for the execution of the crucial metacognitive processes that the learners execute while self-assessing their written products. As we did not assess the processes in which learners engaged while forming their self-assessments, however, we cannot test this potential serial mediation effect via reduced subjective cognitive load, which potentially enhances the degree to which the crucial metacognitive processes could be executed, which, in turn, results in increased judgment accuracy.

It is important to highlight, however, that the benefits of rubrics were not found for relative accuracy and invested mental effort. Regarding the lack of effect on relative accuracy, one explanation could be that the four criteria against which the quality of the abstracts were judged substantially differed in difficulty. Whereas the criteria of content and structure were relatively easy and hence most learners received and provided relatively high
expert and self-assessments, the criteria of sentence structure and grammar and scientific style were relatively difficult, resulting in lower expert and self-assessments. In view of these relatively strong differences, all learners were fairly able to accurately rank their abstracts according to these two pairs of criteria, which overall resulted in relatively high relative accuracy (0.58 in the rubric group and 0.41 in the control group). These values are considerably higher than the moderate value of 0.24 that was found in the recent meta-analysis on relative judgment accuracy in text comprehension by Prinz et al., (2020a; see also Dunlosky & Lipko, 2007). To test this tentative explanation and investigate whether rubrics could in principle increase relative accuracy as well, future studies should include criteria that scarcely differ in difficulty. This should make it harder for learners to accurately rank their performance according to these criteria on their own and thus the scaffolding and offloading functions of rubrics might be more pronounced.

The lack of effect concerning mental effort is not necessarily surprising. Other than the subjective difficulty measure, which indicates the level of difficulty that learners experience during performing a task (here: forming self-assessments), mental effort describes the amount of controlled resources learners allocate to performing a task (van Gog & Paas, 2008; see also Scheiter et al., 2020). Hence, mental effort at least in part reflects learners’ willingness to invest effort and thus perform well on a task (e.g., Roelle & Berthold, 2013; Schnotz et al., 2009). From this perspective, the lack of difference between the groups shows that even for the learners without rubric, the task of forming self-assessments did not reach a level of difficulty that prevented them from investing effort in this task. In view of the findings concerning perceived difficulty and self-assessment accuracy, however, the effort on part of the rubrics learners was better invested than the effort on part of the learners in the no-rubrics group.

Regarding the role of the inferential level of the criteria, our pattern of results indicates that the inferential level does not substantially affect the benefits of rubrics regarding self-assessment accuracy. Hence, Hypothesis 3, which assumed that the potential beneficial effect of rubrics on self-assessment accuracy would decrease as the required inferential level increases, needs to be rejected. At first glance, one explanation for this pattern of results could be that although the learners were only very briefly introduced into the criteria and quality levels, the introduction was nevertheless sufficient for both the low and the high-inferential criteria. At second glance, however, this explanation seems unlikely because in analyzing the potential dependency of the rubric’s benefits on the inferential level of the criteria, we also obtained the finding that across the groups bias was lower for the high inferential than the low inferential criteria. Although this finding does not show that the benefits of rubrics depend on the inferential level of the criteria because this effect occurred in both groups (i.e., regardless of the provision of rubrics), it does show that rubrics could potentially have greater benefits for low inferential than for high inferential criteria, since there is more room for improvement regarding low inferential criteria.

Naturally, these findings concerning the role of the criteria’s inferential level for the effects of rubrics need to be interpreted cautiously before they are replicated and the underlying mechanisms are understood. Furthermore, before the outlined pattern could be related to rubrics in general, it is essential to analyze whether the surprising finding that overall learners’ self-assessments were more accurate regarding high inferential than low inferential criteria is specific for the task of abstract writing that was used in the present study. One potential explanation for the surprising higher self-assessment accuracy on high inferential criteria could be that although the learners had hardly any prior knowledge regarding writing and self-assessing scientific abstracts, the high inferential criteria nevertheless were better aligned with learners’ prior knowledge than the low inferential criteria. More
specifically, as the learners already had classroom instruction on how to write a report, they might to a certain degree already have been able to assess the degree to which a text is written in a concise and clear manner (criterion of sentence structure and grammar) and in precise and objective language (criterion of scientific style), because these criteria apply similarly in different text types. By contrast, the potential prior knowledge on report writing might have been less helpful in self-assessing whether all key aspects of a scientific study were reported (criterion of content) and whether they were reported in the correct order (criterion of structure), because these criteria are specific to the writing of scientific abstracts. To test this tentative explanation, in future studies it would be useful to assess learner prior knowledge regarding the writing of different text types and relate the respective prior knowledge to the accuracy on the high and low inferential criteria.

In these future studies, it could also be fruitful to assess subjective cognitive load in a more fine-grained manner than in the present study. For instance, in order to investigate whether the inferential level of the criteria would affect subjective cognitive load, it could be useful to have learners rate their cognitive load after each self-assessment rather than after forming all self-assessments concerning all criteria. Due to our decision to assess cognitive load only after all self-assessments were formed, we unfortunately cannot say whether the criteria’s inferential level affected cognitive load and whether the above-mentioned effect on subjective difficulty was driven by a reduction of cognitive load during forming the self-assessments concerning the high-inferential criteria in particular.

Limitations and future directions

In addition to the above-mentioned unresolved issues, the present research has some further important limitations. First, it should be mentioned that we did not implement a regulation phase. That is, after the self-assessment phase no revision phase followed. We therefore have no direct evidence that the higher self-assessment accuracy on part of the learners who received rubrics would lead to better task performance. Although theoretical accounts of self-regulated learning support the assumption that higher self-assessment accuracy should foster subsequent regulation (e.g., Nelson & Narens, 1994; see also De Bruin et al., 2020), empirical studies show that increases in accuracy do not always result in effective regulation (see De Bruin & Van Gog, 2012; Roelle et al., 2017). Hence, it would be fruitful if future studies would investigate the potential indirect effects of rubrics on task performance via effects of increased self-assessment accuracy and subsequent regulation. For example, it could be analyzed whether learners who are supported by rubrics in their self-assessment would make better choices concerning the parts of their work that have to be revised, and which instructional support would be needed for the choices to result in increased task performance.

A second important limitation relates to the potential role of task difficulty and task performance. The literature on the related support measure of idea unit standards indicates that the benefits of support measures that are designed to enhance judgment accuracy can depend on learners’ task performance such that the benefits increase with decreasing task performance (see Nederhand et al., 2018). Similarly, Panadero et al. (2016) propose that learners who perform a complex task for the first time usually experience substantial cognitive load while performing the task, which leaves little to no working memory capacity for monitoring their task performance and hence hinders the formation of accurate self-assessments. Hence, it could be assumed that the benefits of rubrics are pronounced for learners who perform poor on the task and hence experience the task as being relatively difficult in particular. Our explorative
moderation analyses did not provide any substantial evidence for this notion. However, our findings concerning the potential moderating role of task performance and difficulty on judgment accuracy and cognitive load in forming self-assessments have to be interpreted very cautiously, because, due to the design of the study, these findings are merely correlational. To adequately address the question whether task performance and/or difficulty would moderate the benefits of rubrics on judgment accuracy and cognitive load in forming self-assessments, future studies would need to experimentally manipulate task performance and/or difficulty (e.g., by implementing support measures while learners perform the task).

A third important limitation relates to the fact that in the present study we provided learners with the rubrics only after they had performed the task. This decision was taken on purpose, because we wanted to outrule the possibility that effects on judgment accuracy would be confounded by effects on task performance. Nevertheless, it could be fruitful to experimentally manipulate the timing of the rubrics in future studies. Arguably, the setting in which learners can work with the rubric already during performing the tasks as well as during revising their products, is more desirable from an applied point of view and likely also more frequently implemented in authentic educational settings. Gaining insight into whether the benefits of rubrics on absolute accuracy, bias, and subjective cognitive load that were observed in the present study would depend on the timing of the rubrics, would thus be highly relevant. Although the present study was based on the notion that previous rubrics research had mainly been conducted in applied settings and would thus entail insufficient internal validity for analyzing the underlying mechanisms of the benefits of rubrics, the very controlled laboratory setting of the present study alone is of course also not the ideal solution. Rather, factors such as the timing of the rubrics that vary between authentic educational settings and more controlled experiments and that could potentially substantially affect the benefits of rubrics should be systematically varied in order to advance the field in future studies.

Finally, it should be highlighted that in the present study the absolute accuracy and bias measures were determined based on four item-response pairs only (i.e., four self-assessments and four expert ratings). As the reliability of both measures can be expected to increase with the number of item-response pairs, at first glance this number seems low and the resultant results as rather unreliable. However, in a wealth of studies that deal with judgment accuracy, absolute accuracy and/or bias are determined on the basis of one item-response pair only (i.e., one overall self-assessment and one overall posttest score, e.g., Kant et al., 2017; Lachner et al., 2020; Prinz et al., 2018). Hence, although a higher number of item-response pairs certainly would have been useful in the present study, the results should nevertheless resonate well in the current self-assessment literature. A similar issue relates to the gamma correlations. Although gamma correlations can be calculated even when only two judgments and objective performances are available (a case Nelson (1984) referred to when he contextualized this measure in research on judgment accuracy), gamma correlations are mostly determined on the basis of at least five judgments per participant (e.g., De Bruin et al., 2011; Thiede et al., 2003). Hence, our results concerning relative accuracy, which did not show any significant effect of the rubric, should be treated cautiously.

Conclusions

Our study entails the following main conclusion: Rubrics beneficially scaffold self-assessments. Evidently, rubrics can increase the accuracy of self-assessments and decrease subjective cognitive load in forming them. Jointly, these findings, which due to the
experimental design and the tightly controlled conditions of the present study entail high internal validity, support the notion that one active ingredient of the established benefits of rubrics on task performance is that rubrics enhance self-assessment accuracy, which should pave the way for effective self-regulation. Notably, these benefits of rubrics were realized without any extensive introduction into how the rubric should be used. Hence, although our findings indicate that more extensive introduction into rubrics might be beneficial, at least certain benefits of rubrics can likely be realized with comparatively economical resources and a low investment of time in educational practice.

**Acknowledgements** We thank the students who participated in our studies and Yu He for her help with the computer programming.

**Author contribution** Conceptualization: Rebecca Krebs, Julian Roelle, Björn Rothstein; Methodology: Rebecca Krebs, Julian Roelle, Björn Rothstein; Formal analysis and investigation: Rebecca Krebs; Writing—original draft preparation: Rebecca Krebs, Julian Roelle; Writing—review and editing: Julian Roelle, Björn Rothstein; Supervision: Julian Roelle, Björn Rothstein.

**Funding** Open Access funding enabled and organized by Projekt DEAL. This work was supported by a grant of the Professional School of Education of Ruhr University Bochum; parts of the study were conducted as part of the doctoral program MeMo-akS of the Professional School of Education.

**Data availability** Not applicable.

**Code availability** Not applicable.

**Declarations**

**Ethics approval** All participants took part on a voluntary basis and their parents 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 parental consent was given for all students under age 18 and own informed consent for all students at the age of 18 or older.

**Consent for publication** not applicable.

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

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.
References

Andrade, H. G. (2001). The effects of instructional rubrics on learning to write. *Current Issues in Education, 4*(4).

Andrade Goodrich, H., & Boulay, B. (2003). Role of rubric-referenced self-assessment in learning to write. *The Journal of Educational Research, 9*(1), 21–30. https://doi.org/10.1080/002202670309596625

Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research, and Evaluation, 10*(1), 3.

Andrade, H., Buff, C., Terry, J., Erano, M., & Paolino, S. (2009). Assessment-driven improvements in middle school students’ writing. *Middle School Journal, 40*(4), 4–12. https://doi.org/10.1080/00940771.2009.11461675

Andrade, H. L., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students’ writing. *Assessment in Education: Principles, Policy & Practice, 17*(2), 199–214. https://doi.org/10.1080/09695941003696172

Ashton, S., & Davies, R. S. (2015). Using scaffolded rubrics to improve peer assessment in a MOOC writing course. *Distance Education, 36*(3), 312–334. https://doi.org/10.1080/01587919.2015.1081733

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

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

Bradford, K. L., Newland, A. C., Rule, A. C., & Montgomery, S. E. (2016). Rubrics as a tool in writing instruction: Effects on the opinion essays of first and second graders. *Early Childhood Education Journal, 44*(5), 463–472. https://doi.org/10.1007/s10643-015-0727-0

Busch-Lauer, I. (2012). Abstracts - eine facettenreiche Textsorte der Wissenschaft [Abstracts - a multifaceted text type of science]. *Linguistik Online, 52*(2). https://doi.org/10.13092/lo.52.293

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

De Bruin, A. B., 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

De Bruin, A. B., & van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction, 22*(4), 245–252. https://doi.org/10.1016/j.learninstruc.2012.01.003

De Bruin, A. B., & van Merriënboer, J. J. (2017). Bridging cognitive load and self-regulated learning research: A complementary approach to contemporary issues in educational research. *Learning and Instruction, 51*(22), 1–9. https://doi.org/10.1016/j.learninstruc.2017.06.001

De Bruin, A. B., Roelle, J., Carpenter, S. K., & Baars, M. (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

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.

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., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students’ learning and retention. *Learning and Instruction, 22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003

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

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*. https://doi.org/10.1007/s11409-022-00929-3

Furze, J., Black, L., Hoffman, J., Barr, J. B., Cochran, T. M., & Jensen, G. M. (2015). Exploration of students’ clinical reasoning development in professional physical therapy education. *Journal of Physical Therapy Education, 29*(3), 22–33. https://doi.org/10.1016/j.hpe.2020.06.002
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

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. Journal of the American Statistical Association, 49, 732–764.

Hayes, A. F. (2013). Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. The Guilford Press.

Hafner, J., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. International Journal of Science Education, 25(12), 1509–1528. https://doi.org/10.1080/0950069022000038268

Hübner, S., Nückles, M., & Renkl, A. (2010). Writing learning journals: Instructional support to overcome learning-strategy deficits. Learning and Instruction, 20(1), 18–29. https://doi.org/10.1016/j.learninstruc.2008.12.001

Jensen, R. (2013). Clinical reasoning during simulation: Comparison of student and faculty ratings. Nurse Education in Practice, 13(1), 23–28. https://doi.org/10.1016/j.nepr.2012.07.001

Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. Assessment and Evaluation in Higher Education, 39(7), 840–852. https://doi.org/10.1080/02602938.2013.875117

Kant, J. M., Scheiter, K., & Oschatz, K. (2017). How to sequence video modeling examples and inquiry tasks to foster scientific reasoning. Learning and Instruction, 52, 46–58. https://doi.org/10.1016/j.learninstruc.2017.04.005

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

Kostons, D., van Gog, T., & Paas, F. (2009). How do I do? Investigating effectsof expertise and performance-process records on self-assessment. Applied Cognitive Psychology, 23, 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., Backfisch, I., Hoogerheide, V., Van Gog, T., & Renkl, A. (2020). Timing matters! Explaining between study phases enhances students’ learning. Journal of Educational Psychology, 112(4), 841–853. https://doi.org/10.1037/edu0000396

Lerdpornkulrat, T., Poondej, C., Koul, R., Khiaiwrod, G., & Prasertsrikul, P. (2019). The positive effect of intrinsic feedback on motivational engagement and self-efficacy in information literacy. Journal of Psychoeducational Assessment, 37(4), 421–434. https://doi.org/10.1177/0734282917744723

Lipko, A. R., Dunlosky, J., Hartwig, M., 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/a0017509

Lipnevich, A. A., McCallen, L. N., Miles, K. P., & Smith, J. K. (2014). Mind the gap! Students’ use of exemplars and detailed rubrics as formative assessment. Instructional Science, 42(4), 539–559. https://doi.org/10.1007/s11251-013-9299-9

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

Murillo-Zamorano, L., & Montanero, M. (2018). Oral presentations in higher education: A comparison of the impact of peer and teacher feedback. Assessment & Evaluation in Higher Education, 43(1), 138–150. https://doi.org/10.1080/02602938.2017.1303032

Nederhand, N., Tabbers, H., Abrahimi, H., & Rikers, R. (2018). Improving calibration over texts by providing standards both with and without idea-units. Journal of Cognitive Psychology, 30(7), 689–700. https://doi.org/10.1080/20445911.2018.1513005

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. Psychological Bulletin, 95(1), 109–133. https://doi.org/10.1037/0033-2909.95.1.109

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition. Metacognition: Knowing about Knowing, 13, 1–25.

Nordrum, L., Evans, K., & Gustafsson, M. (2013). Comparing student learning experiences of in-text commentary and rubric-articulated feedback: Strategies for formative assessment. Assessment & Evaluation in Higher Education, 38(8), 919–940. https://doi.org/10.1080/02602938.2012.758229

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, 1089–1126. https://doi.org/10.1007/s10648-020-09541-1
Rubrics enhance accuracy and reduce cognitive load in…

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., Tapia, J. A., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and Individual Differences, 22*(6), 806–813. https://doi.org/10.1016/j.lindif.2012.04.007

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129–144. https://doi.org/10.1016/j.edurev.2013.01.002

Panadero, E., & Romero, M. (2014). To rubric or not to rubric? The effects of self-assessment on self-regulated learning and self-efficacy. *Assessment in Education: Principles, Policy & Practice, 21*(2), 133–148. https://doi.org/10.1080/13603124.2013.1016784

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22, 74–98*. https://doi.org/10.1016/j.edurev.2017.08.004

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.

Panadero, E., & Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. *Educational Research Review, 30*, 100329. https://doi.org/10.1016/j.edurev.2020.100329

Prinz, A., Golke, S., & Wittwer, J. (2018). The double curse of misconceptions: Misconceptions impair not only text comprehension but also metacomprehension in the domain of statistics. *Instructional Science, 46, 723–765*. https://doi.org/10.1007/s11251-018-9452-6

Prinz, A., Golke, S., & Wittwer, J. (2020a). 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

Prinz, A., Golke, S., & Wittwer, J. (2020b). To what extent do situation-model-approach interventions improve relative metacomprehension accuracy? Meta-analytic insights. *Educational Psychology Review, 32*, 917–949. https://doi.org/10.1007/s10648-020-09558-6

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics 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

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435–448. https://doi.org/10.1080/02602930902862859

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/10.1111/cogs.12086

Roelle, J., Krüger, S., Jansen, C., & Berthold, K. (2012). The use of solved example problems for fostering strategies of self-regulated learning in journal writing. *Education Research International*. https://doi.org/10.1155/2012/751625

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*, 99–117. https://doi.org/10.1037/edu0000132

Rost, D. H., Wirthwein, L., Frey, K., & Becker, E. (2010). Steigert Kaugummikauen das kognitive Leistungsvermögen? [Does chewing gum enhance cognitive abilities? Two rather extraordinary experiments?]. *Zeitschrift Für Pädagogische Psychologie, 24*(1), 39–49. https://doi.org/10.1024/1010-0652.a000003

Sadler, P. M., & Good, E. (2006). The impact of self-and peer-grading on student learning. *Educational Assessment, 11*(1), 1–31. https://doi.org/10.1207/s15326977ea1101_1

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

© Springer
Schamber, J. F., & Mahoney, S. L. (2006). Assessing and improving the quality of group critical thinking exhibited in the final projects of collaborative learning groups. *The Journal of General Education, 55*(2), 103–137. https://doi.org/10.1353/jge.2006.0025

Scheiter, K., Ackerman, R., & Hoogerheide, V. (2020). Looking at mental effort appraisals through a metacognitive lens: Are they biased? *Educational Psychology Review, 32*, 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

Schnitz, W., Fries, S., & Horz, H. (2009). Motivational aspects of cognitive load theory. In M. Wosnitza, S. A. Karabenick, A. Efklides, & P. Nenniger (Eds.), *Contemporary motivation research: From global to local perspectives* (pp. 69–96). Hogrefe & Huber Publishers.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and learning, 4*(1), 33–45.

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

Smit, R., Bachmann, P., Blum, V., Birri, T., & Hess, K. (2017). Effects of a rubric for mathematical reasoning on teaching and learning in primary school. *Instructional Science, 45*(5), 603–622. https://doi.org/10.1007/s11251-017-9416-2

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Swales, J. M. & Feak, C. B. (2009). *Abstracts and the writing of abstracts* (Vol. 2). University of Michigan Press ELT.

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., & 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

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., & 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

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

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.