Metacognition and Learning (2026) 21:44
https://doi.org/10.1007/s11409-026-09492-y

The role of performance level descriptions for the effects of rubrics on self-assessment accuracy and performance

Rebecca Krebs[1] · **Julia Waldeyer**[2] · **Björn Rothstein**[3] · **Ernesto Panadero**[4,5,6] · **Julian Roelle**[7]

Received: 10 November 2025 / Accepted: 12 August 2026
© The Author(s) 2026, modified publication 2026

Abstract
Rubrics are widely used to enhance self-regulated learning and academic performance, yet the specific roles of their main components (i.e., assessment criteria and performance level descriptions) remain unclear. Using the case of a writing task (i.e., writing an abstract for a scientific article), the present experiment (N = 236) employed a 2×2 factorial design to isolate the effects of the component of performance level descriptions on self-assessment accuracy after an initial writing phase and on performance gains in a revision phase. Results showed that performance level descriptions significantly improved learners' self-assessment accuracy and entailed differential effects on revision performance. When provided before the revision phase (i.e., during self-assessment), performance level descriptions fostered effective revision, but only for learners who already performed relatively well in the initial writing phase. For learners who performed poorly in the initial writing phase, by contrast, performance level descriptions during self-assessment did not foster – and in some analyses even hindered – subsequent revision. When provided during revision, performance level descriptions scarcely affected performance. The findings underscore the importance of performance level descriptions for accurate self-assessment but not necessarily for task performance, offering guidance for optimizing rubric design in educational practice.

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

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

2 Institute of Educational Psychology, Technische Universität Braunschweig, Braunschweig, Germany

3 Faculty of Philology, Ruhr University Bochum, Bochum, Germany

4 Institute of Education, St. Patrick's Campus, Dublin City University, Dublin, Ireland

5 Education Regulated Learning and Assessment (ERLA group), Facultad de Educación y Deportes, Universidad de Deusto, Bilbao, Spain

6 IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

7 Department of Psychology, University of Münster, Münster, Germany

Published online: 04 September 2026

Using rubrics to communicate assessment criteria and performance level descriptions is a common instructional strategy around the world (e.g., Brookhart, 2018; Panadero et al., 2023). And that is a good thing—a recent meta-analysis shows that rubrics foster different facets of self-regulated learning (e.g., the accuracy of learners' self-assessments) and performance across various academic tasks and student populations (Panadero et al., 2023). Despite these promising findings, however, there are some important open questions concerning the effects of rubrics that still need to be addressed.

One important open question concerns the specific roles of the two main components of rubrics: assessment criteria and performance level descriptions. Most previous studies have compared rubrics to control conditions that included neither of these components (e.g., Andrade et al., 2009; Panadero et al., 2012). As a result, it is still unclear how much each component contributes to the positive effects of rubrics.

The present study was designed to explore this open question by focusing on the role of *performance level descriptions*. From a theoretical perspective, the added value of this component could depend on the specific function rubrics are intended to serve. When the goal is to support learners in making accurate self-assessments, we expect that a full rubric (including assessment criteria *and* performance level descriptions) will be more effective than only providing learners with the assessment criteria alongside a scale that differentiates performance levels but does not describe them. However, when the goal is to support performance improvement, providing learners with the assessment criteria alongside a scale that differentiates but does not describe different performance levels (in the following referred to as *rating scale*) may be sufficient.

To test these assumptions, we conducted a 2×2 factorial experiment. We varied whether learners received full rubrics or rating scales during two key phases: (1) a self-assessment phase after an initial phase of working on a writing task, and (2) a subsequent revision phase in which learners were to improve their performance on the writing task (the rubrics and rating scales were the same in both phases). The main outcome measures were self-assessment accuracy and performance gains in the revision phase.

Rubrics effects on self-assessment accuracy and task performance

Rubrics provide learners with assessment criteria (e.g., in assessing the quality of an abstract for a scientific article, such criteria could be whether the main research questions and the key results are mentioned) and, based on these assessment criteria, define performance levels for academic tasks (e.g., which criteria need to be met to what degree for an excellent, a good, or a satisfactory performance), typically arranged along a continuum (e.g., Brookhart, 2018; Panadero & Jonsson, 2013). These criteria and performance levels are usually organized in the form of a table or matrix (Panadero et al., 2023) and can relate to only one or to several different assessment dimensions. However, it has scarcely been explored whether both main components of rubrics (i.e., assessment criteria *and* performance level descriptions) are necessary for these effects to occur.

Previous research indicates that providing learners with assessment criteria alone can be beneficial. For example, Fastré et al. (2010) found that the provision of assessment criteria improved learner performance in a test phase. Similarly, Balan and Jönsson (2018) obtained evidence that at least partially showed that the provision of assessment criteria fostered

learners' performance as well as their self-efficacy. In view of these findings, the question arises whether providing learners with "full" rubrics is actually necessary and hence has added value to the provision of assessment criteria. As Brookhart (2018) and Brookhart and Chen (2015) describe, performance level descriptions are the main component that differentiates mere assessment criteria from rubrics. Unfortunately, most previous studies either compared rubrics' use to conditions that did not receive any information regarding assessment criteria and/ or performance levels (e.g., Andrade et al., 2009; Kocakülah, 2010; Lerdpornkulrat et al., 2019), or the studies did not differentiate between the provision of mere assessment criteria (without performance level descriptions) and the provision of rubrics (with performance level descriptions), which makes it unclear whether solely providing assessment criteria would already explain the beneficial effects of rubrics. In the present study, we focus on the role of the component of performance level descriptions. From a theoretical perspective, it is reasonable to assume that this component is important for the effects on self-assessment accuracy but not for the effects on task performance.

Performance level description effects on self-assessment accuracy

The theoretical rationale for the positive effects of rubrics on self-assessment accuracy is grounded in the cue-utilization framework (Koriat, 1997). According to this framework, learners form self-assessments, such as judgments about their understanding or performance, based on cues they can monitor. These cues may include, for example, response length (e.g., the length of a written abstract for a scientific study), interest in the topic, or perceived mental effort during performing a task (e.g., Baars et al., 2020; Froese & Roelle, 2022; Thiede et al., 2010; Waldeyer et al., 2024; for relations between self-assessments and mental effort see also De Bruin et al., 2020). The accuracy of self-assessments depends on how diagnostic these cues are, that is, how well they predict actual performance (De Bruin et al., 2017, 2020).

Rubrics are assumed to enhance self-assessment accuracy by helping learners attend to more diagnostic cues. In support of this notion, Krebs and colleagues (2022, 2024) found that rubrics increased absolute accuracy of learners' self-assessments and decreased bias, which might be due to the guidance rubrics provide in attending to diagnostic cues. Notably, although they changed the way learners formed self-assessments, the rubrics did not increase (but partly even decreased) cognitive load during self-assessment. Likely, the assessment criteria that were communicated by the rubrics directly guided learners in monitoring diagnostic cues. However, the assessment criteria may not be the sole driving force behind the benefits of rubrics regarding self-assessment accuracy. Learners can evaluate the extent to which they meet each assessment criterion, but without *performance level descriptions*, they may struggle to interpret the quality of their work as a whole. These descriptions provide benchmarks that help translate evaluations of individual assessment criteria into overall performance judgments. Thus, performance level descriptions may enhance self-assessment accuracy not by introducing new cues, but by enabling learners to make better use of the cues they already have. Despite this theoretical claim, direct empirical tests of this hypothesis remain limited. Importantly, testing this hypothesis requires assessing both the cues learners rely on and the ways in which they use them in forming self-assessments. One approach to do so is to ask learners to think aloud in forming self-assessments (Bol et al., 2010; Dinsmore & Parkinson, 2013).

Performance level description effects on task performance

Unlike for the outlined potential beneficial effects on self-assessment accuracy, the role of performance level descriptions for task performance remains theoretically contested. On the one hand, such descriptions might help learners monitor their progress more effectively. By clarifying what different levels of quality look like, performance level descriptions could support formative self-assessment during working on a task or revising one's task performance (e.g., in revising a written summary or abstract for a scientific article), helping learners judge whether they have reached a satisfactory level and identify how far they are from a higher one. In doing so, these descriptions might enable learners to allocate their efforts more efficiently, focusing on aspects most in need of improvement and thereby foster performance gains.

On the other hand, it is plausible that performance level descriptions do not offer substantial added value in this context. Learners might be able to enhance their task performance effectively simply by considering how many of the assessment criteria they have met and trying to meet further ones, without needing to position themselves along a quality continuum. In fact, introducing detailed descriptions of performance levels might impose additional cognitive demands during performing a task (e.g., during revising one's work). Learners would likely process this information even if it does not directly inform their actions, thereby potentially increasing extraneous cognitive load. This, in turn, could detract from the cognitive resources available for executing meaningful revisions as well as reduce instructional efficiency as measured by performance gains in relation to experienced cognitive load (Roelle & Berthold, 2013; Van Gog & Paas, 2008). Empirical studies that directly test these competing assumptions, however, are still scarce.

Performance level descriptions may, however, support task performance indirectly when provided during a self-assessment phase before learners are given the opportunity to revise their work. In tasks involving multiple assessment dimensions and under constraints of time or cognitive capacity (which are common in authentic educational settings), helping learners generate more accurate self-assessments can be crucial. When learners are supported with performance level descriptions during self-assessment, they may be better able to identify specific weaknesses and prioritize what to revise. This idea is supported by research showing that learners who receive instructional support for self-assessment tend to benefit more from subsequent opportunities to revise knowledge gaps (Thiede et al., 2003, 2017). Accordingly, although performance level descriptions might have little effect *during* performing a task itself, they could still pave the way for effective revision and hence contribute to increases of task performance when provided earlier in the learning cycle.

Aim, research questions, and hypotheses

In view of the outlined considerations, our main aim was to investigate the effects of rubric performance level descriptions on self-assessment accuracy and task performance. For this purpose, we compared rubrics (which include both assessment criteria and performance level descriptions) against rating scales (which include assessment criteria but do not contain performance level descriptions). We formulated two research questions and their corresponding hypotheses.

RQ 1. What is the effect of performance level descriptions on self-assessment accuracy?

We hypothesized that providing learners with rubrics would enhance self-assessment accuracy in comparison to providing them with rating scales only (Hypothesis 1a). Furthermore, we predicted that the assumed potential beneficial effect of rubrics would be mediated through the rubrics enabling learners to establish more meaningful relations (i.e., relations of better quality) between the degree to which the assessment criteria are met and different performance levels (Hypothesis 1b). These relations were assessed through think-aloud methodology.

RQ 2. What are the effects of performance level descriptions on task performance and instructional efficiency in a revision phase?

We hypothesized that providing learners with rubrics rather than rating scales *before* they perform a task (in the present study: before learners revise their initial task performance) would lead to higher performance gains in a revision phase (Hypothesis 2a). However, when rubrics are provided during revision, we hypothesized that they would not foster performance gains in comparison to providing learners with rating scales (Hypothesis 2b). Furthermore, we assumed that instructional efficiency (i.e., performance gains in relation to experienced cognitive load) would be higher when only rating scales were provided during revision (Hypothesis 2c).

Method

Design and sample

The experiment followed a 2 × 2 factorial between-subjects design where the factors *rubric vs. rating scale during self-assessment* and *rubric vs. rating scale during revision* were varied, which resulted in four conditions. All learners read a scientific article and were instructed to write an abstract for this article. At the end of the writing phase, learners were asked to self-assess the quality of their abstracts on six assessment dimensions. During self-assessment, based on random assignment, the learners received either the rubrics or the rating scales. Then, learners decided on which of the six assessment dimensions they would like to receive further instruction (i.e., learners took regulation decisions). They could choose as many dimensions as they wanted to. Afterwards, they were to revise their abstracts on the chosen dimensions. During revision, based on random assignment, we again varied whether learners were supported by rubrics or only received the rating scales.

We determined the required sample size for the experiment by running an a priori power analysis with G*Power (Faul et al., 2007). Based on the minimum desired effect of instructional interventions of $d = 0.40$ according to Hattie (2009), we used a targeted effect size of $d = 0.40/f = 0.20$ in this analysis. The further parameters were $\alpha = 0.05$, and $\beta = 0.20$ (i.e., power of 80%). In a 2 × 2 factorial between-subjects design, the power analysis indicated that a sample of 199 participants would be required to detect the effects at which our research questions were targeted. We slightly oversampled and recruited a total of $N = 236$ German university students (186 female; $M_{Age} = 22.71$ years, $SD_{Age} = 2.77$ years). University students were bachelor's and master's students from different subjects. Bachelor students accounted for 48.3% of the sample, master students for around 51.7%. The majority of participants was enrolled in Humanities or Social Science programs which also included preservice teachers (87.3%), followed by students from the domain of Law or Architecture (5.1%), Natural Sci-

ences (3.8%), Technology and Engineering (3.0%), and Economics (0.8%). The voluntary participation was compensated with €20 and written informed consent was collected from all participants. This study's design and the analyses were not pre-registered. The study was approved by the ethics committee of Ruhr University Bochum (EPE-2022-35).

Materials

All materials were provided in a computer-based learning environment. The materials were based on the material used by Krebs et al. (2022). All materials were in German and an expert in the field of German didactics (i.e., the third author of this study) supervised the entire material development.

Introduction to writing a scientific abstract

All learners received an introduction in how to write a scientific abstract. This introduction was provided in written form with short explanations and bullet points. First, learners were informed about what an abstract is. Based on Swales and Swales (1990), Swales and Feak (2009), and Busch-Lauer (2012), the learners were informed that abstracts are short summaries of a scientific article and should give the reader an overview about the empirical study that is presented in the article. Moreover, they were instructed that an abstract should contain the most important aspects of the study (for empirical research: research questions or hypotheses, the data basis, method, results, and discussion).

Afterwards, they received information about how to write a scientific abstract. This information was structured around six assessment dimensions, which were the same dimensions on which learners were to self-assess their performance later: (a) content, (b) formal, (c) orthography and grammar, (d) scientific style, (e) text organization, and (f) structure. For each dimension, learners were informed about the important assessment criteria that needed to be met (these criteria were also part of the rubrics and rating scales, see section *Rubrics and Rating Scales During Self-Assessment and Revision*). For instance, for the dimension of (a) content, the learners were informed that research questions/hypotheses, the data basis, the method, the results, and a brief conclusion of the results needed to be included.

To help learners discern and understand the six dimensions, they received contrasted exemplars (one of high quality and one of low quality) for each dimension. These exemplars illustrated the assessment criteria of the dimensions, but their content was not related to the content of the scientific article.

Scientific article and writing task

All participants read a slightly adapted scientific article that was published under the title *"Steigert Kaugummikauen das kognitive Leistungsvermögen? Zwei Experimente der besonderen Art"* [*Does Chewing Gum Enhance Cognitive Abilities? Two Rather Extraordinary Experiments*] by Rost et al. (2010). The article consisted of 1,810 words and was fitting the criteria of a well-written scientific article. It was slightly adapted so that it was understandable without prior knowledge on statistics. The learners were asked to read the article and afterwards were instructed to write an abstract for it. Learners were prompted to keep in mind what they had learned in the introduction about the criteria that an abstract should

meet. During writing, which was limited to 20 min, the article was available and a timer showing the remaining time was displayed.

Rubrics and rating scales during self-assessment and revision

After the writing phase, all learners were required to self-assess the quality of their abstracts. Learners in the conditions who received the support of rubrics during self-assessment were given a specific rubric for every dimension (i.e., six in total) separately, and these were displayed to them one after another. Each rubric contained the two main components a rubric is made of, that is, assessment criteria and performance level descriptions as well as a rating scheme (from 1: very poor quality, to 6: very good quality). Hence, the learners were to provide one self-assessment for each of the six assessment dimensions (resulting in six self-assessment scores). The learners with only the rating scales during self-assessment also self-assessed their performance on the six dimensions.

However, in contrast to the conditions that were supported by rubrics, they were presented only one of the main components of rubrics, that is, they only received the assessment criteria and a rating scheme (which was the same as for learners with rubrics, ranging from 1 as very poor quality to 6 as very good quality). Similar to the learners with rubrics, they were asked to assess themselves on each dimension separately, resulting in six self-assessment scores as well. Figure 1 shows the self-assessment phase for learners with rubrics vs. for learners with rating scales for the dimension of text organization. Screenshots from the self-assessment phase in the learning environment for all dimensions can be found in Appendix A. During self-assessment, in all conditions learners could see their self-written abstracts but could not modify them.

Following the self-assessments, learners were to take regulation decisions. Specifically, all learners were instructed: "Now it's time to improve your abstract. You now have the option of displaying further exemplars of the dimensions you have already learned about and then revising your abstract. You can choose between zero and all six dimensions. Please note: the number of dimensions you select does not affect the duration of the study; the study will always be the same length.[1] Please now click on all dimensions for which you would like to receive further information and which you would like to improve." Based on

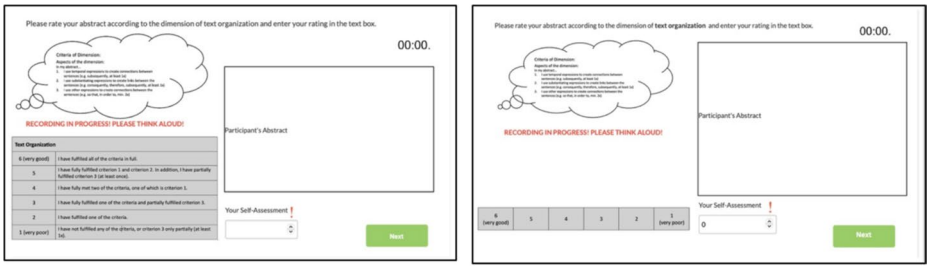


Fig. 1 Self-assessment with rubric vs. with rating scale for the dimension of text organization (translated from German)

[1] This information was given to ensure that learners would not choose fewer dimensions than necessary in order to quit the study as early as possible. However, contrary to this information the study ended when the individual learners had finished their abstracts. The learners were informed about this deception at the end of the experiment.

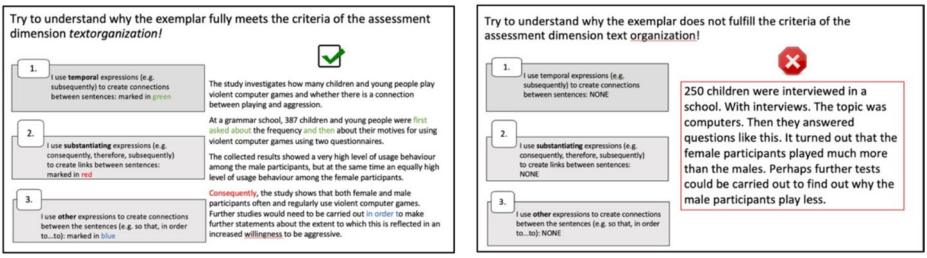


Fig. 2 Exemplars for the dimension of text organization (high and low quality) that were shown to the participants after the regulation decision (translated from German)

their regulation decision(s), the learners received two exemplars for each chosen dimension (one of high and one of low quality) and were instructed: "Try to understand why the exemplar fully meets/does not fulfill the assessment criteria of the dimension *content/formal/orthography and grammar/scientific style/text organization/structure*!" The exemplars for the dimension of text organization are shown in Fig. 2.

Immediately after studying the two exemplars for a dimension, the learners were asked to revise their abstract concerning the respective dimension with the following prompt: "Please revise your abstract here with regard to the dimension *content/formal/orthography and grammar/scientific style/text organization/structure!"* During this revision phase, we varied whether learners were supported with rubrics or only received the rating scales (the rubrics and rating scales were identical to the ones used during self-assessment). There was no second self-assessment phase at the end of the revision phase.

Instruments and measures

Task performance

We measured task performance at two time points, that is, after the initial writing phase and after the revision phase. For this purpose, two trained raters assessed the quality on each of the six assessment dimensions using the same rubric the learners received. Between 1 (very poor) and 6 (very good) points could be awarded for each dimension. Interrater reliability, determined by the intraclass correlation coefficient with measures of absolute agreement, was very good (for all dimensions, ICC > 0.85). The scores of all six dimensions were aggregated, resulting in a minimum of 6 points and a maximum of 36 points as an overall performance score.

Self-assessment accuracy

To compute the accuracy of learners' self-assessments, we compared the six assessments of the learners with the assessments of the experts after the initial writing phase and calculated three different measures. First, we calculated the absolute discrepancy between learners' and experts' ratings ($|X_{\text{learners' assessment}} - X_{\text{experts' assessment}}|$), that is, *absolute accuracy* (see Schraw, 2009). This measure describes the degree of correspondence between two ratings and allows values between 0 (i.e., perfect accuracy) and 5 (i.e., max. inaccuracy) for each dimension. We aggregated these scores by summing them up over the six dimensions (i.e., scores between 0 and 30). Second, we calculated *bias*, which is the signed difference

between the learners' assessments and the experts' ratings, allowing to account for over- or underconfidence. Bias was calculated by first subtracting the experts' assessment from the learners' assessment ($X_{\text{learners' assessment}}$- $X_{\text{experts' assessment}}$) separately for each dimension (six scores between −5 and 5) and averaging the six scores afterwards. As a result, we receive one value for each learner that varied between −5 and 5. Here, positive values indicate that the learner overestimated their performance, values close to zero indicate high accuracy, and negative values signal underconfidence.

Think aloud

In order to assess the processes in which the learners engaged during self-assessment and hence to be able to test our hypothesis that learners with rubrics would establish more meaningful relations (i.e., relations of better quality) between the degree to which the assessment criteria are met and different performance levels (Hypothesis 1b), learners were instructed to think aloud. Specifically, after a brief introduction to the think-aloud methodology, based on recommendations by Ericsson and Simon (1993) the learners were prompted: "[…] Please say out loud everything that comes to your mind and goes through your head while you are solving the task. It is important that you do not try to explain or structure what you are doing. Just imagine that you are alone in the room and talking to yourself. […]".[2]

These think-aloud protocols were analyzed using the following coding procedure. First, as a proxy for whether learners attended to predictive cues, we determined the number of the assessment criteria (these were included in both the rubrics and the rating scales) that learners mentioned. Second, we determined how learners related the degree to which the assessment criteria of a dimension were met to the respective performance levels by rating their quality using a 3-point rating scale. As learners in both conditions were provided with a rating scheme that included six performance levels, these relations could be established by all learners. However, only the learners who received rubrics during self-assessment also received performance level descriptions and hence establishing meaningful relations between the assessment criteria and performance levels was deemed more likely for these learners to occur. Specifically, the quality of these relations to performance levels was rated as high (i.e., a score of 2) if the self-assessment processes were justified in a differentiated manner, that is, performance level descriptions or scale points were mentioned and set in relation to the self-written abstract and/or further differentiated explanations were given to why this assessment was made. The quality of relations to performance levels was rated as medium (i.e., a score of 1) if the self-assessment processes were rudimentarily justified, that is, the performance level descriptions/scale points were mentioned but not further elaborated, and no further explanations were given as to why this assessment was made. A low-quality rating of relations to performance levels (i.e., a score of 0) was provided when self-assessment processes were not justified, that is, performance level descriptions or scale points were not mentioned or differentiated, the criteria were not weighted against each other, and no further explanation was given as to why the specific assessment was made.

The quality scores were determined for the six dimensions separately and aggregated into an overall score. Interrater reliability was very good for counting the utilization of predictive cues (ICC > 0.94) and for rating the quality of the relations to performance levels (ICC > 0.83).

[2] The learners were also instructed to think aloud while taking their regulation decisions. However, as we did not have a specific research question or hypothesis regarding this process, we do not report this data in the present manuscript.

Cognitive load

We measured learners' subjective cognitive load at five different time points during the experiment: after reading the scientific article, after writing their abstract, after self-assessment, after the regulation decisions and after the revision phase. Specifically, on each measurement time point, we asked the learners to rate their perceived difficulty and mental effort on 7-point Likert scales (*1: very low difficulty/mental effort*, *7: very high difficulty/mental effort*). The wording of the questions was adapted from previously existing scales for the assessment of mental effort and perceived difficulty in task processing (Kalyuga et al., 1999; Paas, 1992; Paas et al., 2003).

Instructional efficiency

Instructional efficiency was measured according to the measure introduced by Paas and Van Merriënboer (1993), which combines measures of test or task performance with measures of subjective cognitive load required to attain this performance (see also Van Gog & Paas, 2008). Specifically, we used the improvements of task performance (i.e., performance gains) the learners realized during the revision phase, standardized them and combined them with learners' ratings of perceived difficulty and mental effort (these measures were averaged) during the revision phase according to the formula suggested by Paas and Van Merriënboer (1993), $\frac{zP_{Improvement} - zE_{Cognitive\ Load}}{\sqrt{2}}$.

Procedure

All learners worked individually in a digital learning environment in a laboratory in soundproofed cubicles to avoid distractions. All learners were asked to give general demographic information on gender, age, GPA score and university semester. Then, all learners received the thematic introduction in how to write a scientific abstract, which included the different dimensions and assessment criteria. Afterwards, they read the scientific article. Reading time was held constant at 20 min. Next, they wrote their abstract for the article. Writing time was held constant at 20 min.

In the next step, the experimental manipulation took place. Learners were randomly assigned to one of the four conditions and were asked to self-assess their abstracts concerning the six dimensions, either with the help of rubrics or rating scales. We prompted all learners to think aloud while forming their self-assessments. Following the self-assessments, the learners took their regulation decisions and were prompted to think aloud while doing so. Once they had taken their regulation decisions, learners processed the exemplars of the chosen dimensions and were to revise their abstracts (this was done separately for each dimension, although learners could in principle implement revisions that affected other dimensions as well), either with the help of rubrics or rating scales, in a self-paced manner. In this phase, rubrics and ratings scales were only implemented, but no additional explicit assessment was required from the learners. After completing their revisions for the selected dimensions one by one, the experiment ended. The experiment took approximately 1.5 h. A detailed overview of the procedure can be found in Fig. 3.

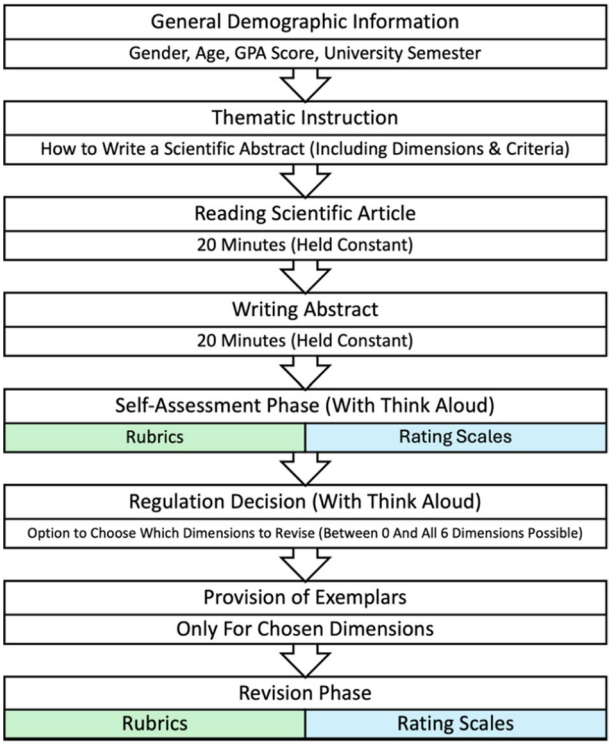


Fig. 3 Overview of the procedure

Results

An α-level of 0.05 was used for all analyses (except the analyses concerning equivalence, see Hypothesis 2b). Tables 1 and 2 show the mean scores and standard deviations for all groups on all measures of the experiment. Data and analyses codes can be found under: https://doi.org/10.17605/OSF.IO/7ZGVR.

Data analyses

All analyses were performed with IBM SPSS Statistics, Version 29.0. To test for potential a priori differences between the experimental conditions, we conducted ANOVAs and Chi-square tests. To answer our hypotheses regarding the effects on self-assessment accuracy (Research Question 1), we used *t* tests as well as a mixed-repeated measures ANOVA to account for potential differences between the dimensions (Hypothesis 1a).

Additionally, to investigate the potential mechanism via which the performance level descriptions affect self-assessment accuracy, we conducted a mediation analysis with 95% percentile bootstrap confidence intervals from 10,000 bootstrap samples using the SPSS macro PROCESS 4.0 (Hayes, 2022; see Hypothesis 1b).

To test our hypotheses concerning the effects on performance gains in the revision phase (Research Question 2), we again used ANOVAs and equivalence testing to account for the predicted non-significance of results regarding Hypothesis 2b.

Table 1 Means and standard deviations of all variables used in the preliminary analyses

	Rubrics – Rubrics (n = 58; 45 female)		Rubrics – Rating Scales (n = 61; 49 female)		Rating Scales – Rubrics (n = 58; 47 female)		Rating Scales – Rating Scales (n = 59; 45 female)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Grade point average (1–6)	2.07	0.58	2.09	0.53	2.13	0.56	2.02	0.52
Age (18–30)	22.76	2.64	22.48	2.72	22.62	3.13	22.98	2.64
Task difficulty (reading) (1–7)	2.71	1.23	3.08	1.44	2.98	1.07	3.19	1.47
Mental effort (reading) (1–7)	4.26	1.36	4.38	1.32	4.78	1.38	4.34	1.41
Task difficulty (writing) (1–7)	3.49	1.33	3.85	1.33	3.81	1.39	3.83	1.31
Mental effort (writing) (1–7)	4.47	1.23	4.64	1.45	4.90	1.22	4.81	1.23
Task performance after initial writing phase (6–36)	28.43	3.78	28.37	4.24	29.29	3.28	28.34	4.62

Table 2 Means and standard deviations of all dependent variables

	Rubrics – Rubrics (n=58; 45 female)		Rubrics – Rating Scales (n=61; 49 female)		Rating Scales – Rubrics (n=58; 47 female)		Rating Scales – Rating Scales (n=59; 45 female)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Self-assessment: Absolute accuracy (0–30)	5.79	3.08	5.29	2.47	6.24	3.63	7.05	3.30
Self-assessment: Bias (−5–5)	0.10	0.68	0.03	0.47	−0.14	0.75	−0.01	0.87
Self-assessment: Use of predictive cues (0–∞)	19.42	5.95	21.43	2.19	20.72	3.12	20.28	3.29
Self-assessment: Quality of references to performance levels (0–2)	1.02	0.59	1.19	0.49	0.51	0.41	0.45	0.40
Self-assessment: Task difficulty (1–7)	2.81	1.09	2.93	0.95	2.80	0.82	2.74	0.89
Self-assessment: Mental effort (1–7)	3.04	1.25	3.13	1.18	3.52	1.34	3.02	1.10
Regulation: Task difficulty (1–7)	2.93	1.43	2.93	1.26	2.75	1.40	2.96	1.56
Regulation: Mental effort (1–7)	2.88	1.58	3.10	1.48	3.25	1.68	3.12	1.60
Revision: Task performance (6–36)	30.28	3.33	29.98	4.54	30.90	3.17	30.05	3.66
Revision: Task difficulty (1–7)	2.70	1.41	3.14	1.36	2.87	1.23	2.96	1.35
Revision: Mental effort (1–7)	2.95	1.43	3.33	1.40	3.79	1.67	3.25	1.30
Revision: Instructional efficiency	0.23	0.86	−0.07	0.99	−0.14	0.96	0.04	1.24
Revision: Performance gains (−30–30)	1.84	2.29	1.61	2.07	1.60	2.47	1.71	3.06

Due to technical errors, data was partially not recorded, which is the reason for the slightly different degrees of freedom in the analyses. As the missing data related to only five participants (i.e., ca. 2% of the sample), we did not try to impute the missing data points. We did not exclude any outliers from our analyses.

Preliminary analyses

In the first step, we tested whether the random assignment resulted in groups that were comparable in terms of learning prerequisites. ANOVAs did not reveal any statistically significant differences concerning age, $F(3, 232)=0.36$, $p = .783$, $\eta_p^2<0.01$, or GPA, $F(3, 232)=0.43$, $p = .735$, $\eta_p^2<0.01$. Also, the groups did not significantly differ with respect to gender, $\chi^2(3)=0.54$, $p = .911$. Furthermore, the groups did not differ in terms of study domain, $\chi^2(12)=14.96$, $p = .244$ or in the intended degree, $\chi^2(6)=7.97$, $p = .240$, what suggests an even distribution across all groups.

In the second step, we tested whether there were any differences between the groups concerning the reading phase and the initial phase of writing the abstract, which both took place before the experimental manipulations were implemented. In the reading phase, there were no statistically significant differences concerning both measures of subjective cognitive load, $F(3, 230)=1.45$, $p = .228$, $\eta_p^2=0.02$ for perceived difficulty, and $F(3, 230)=1.61$, $p = .188$, $\eta_p^2=0.02$ for invested mental effort. Likewise, there were no statistically significant differences concerning cognitive load in the initial phase of writing the abstract, $F(3, 230)=0.93$, $p = .426$, $\eta_p^2=0.01$ for perceived difficulty, and $F(3, 230)=1.22$, $p = .303$, $\eta_p^2=0.02$ for invested mental effort. Also, the groups did not differ concerning task performance in the initial writing phase (i.e., the overall quality of their abstracts at the end of the initial writing phase), $F(3, 232)=0.76$, $p = .520$, $\eta_p^2=0.01$. Jointly, these results indicate that the random assignment resulted in comparable groups.

RQ 1: What is the effect of performance level descriptions on self-assessment accuracy?

In Hypothesis 1a, we assumed that providing learners with rubrics (i.e., assessment criteria, rating scheme, *and* performance level descriptions) would enhance self-assessment accuracy in comparison to providing them with only rating scales (i.e., assessment criteria and a rating scheme). We found a statistically significant difference between the learners with rubrics and the learners with rating scales in terms of absolute accuracy, $t(234) = 2.71$, $p = .004$, $d = 0.35$ (one-sided). As expected, the learners with rubrics showed better absolute accuracy (i.e., values closer to zero) than their counterparts ($M_{\text{rubrics}} = 5.54$, $SD = 2.78$; $M_{\text{rating scales}} = 6.65$, $SD = 3.48$).[3]

In terms of bias, we did not find a statistically significant difference between the learners with rubrics and with rating scales, $t(210.23) = -1.52$, $p = .065$, $d = -0.20$ (one-sided). On a descriptive basis, learners with rubrics showed a slight overconfidence ($M_{rubrics} = 0.07$, $SD = 0.58$) whereas learners with rating scales showed underconfident self-assessments ($M_{rating\ scales} = -0.07$, $SD = 0.81$).

In Hypothesis 1b, we assumed that the potential superiority of the rubrics over the rating scales concerning self-assessment accuracy would be mediated through the rubrics enabling learners to establish more meaningful relations (i.e., through their quality relations) between the degree to which the assessment criteria were met and the performance levels. The path results of the mediation analysis regarding the effect on absolute accuracy are shown in Fig. 4. We found a statistically significant indirect effect via the quality of the established relations between the degree of how the criteria were met and different performance levels, $a \times b = -0.54$ [−0.76, −0.32]. This result suggests that due to the mediating function of the quality of the established relations, the learners with rubrics had an advantage of −0.54 units in terms of absolute accuracy, meaning that with an increase of one unit in quality, learners with rubrics had a 0.54 units more accurate self-assessments than learners with only the rating scales. Given that the effect predicted in Hypothesis 1a was significant only for absolute accuracy, we refrained from testing Hypothesis 1b using bias as the outcome measure.

For exploratory purposes, we also analyzed whether supporting learners with rubrics rather than with rating scales during self-assessment would affect cognitive load in forming self-assessments. However, there were no statistically significant differences between

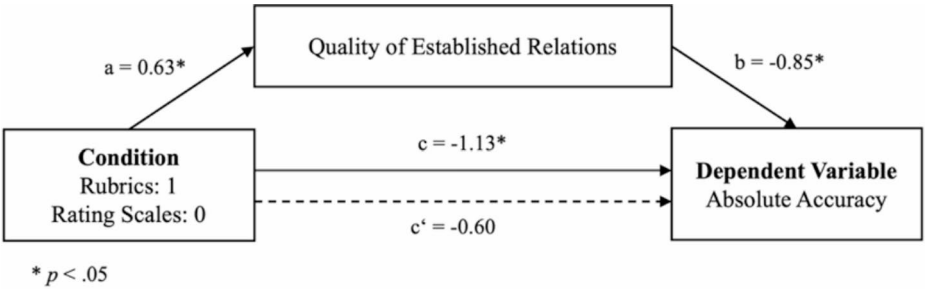


Fig. 4 Results of the mediation analysis concerning the effect on absolute accuracy

[3] For explorative purposes, we tested whether this effect was moderated by learners' performance in the initial writing phase. However, there was no significant moderation effect, $p = .305$, $\eta_p^2 < 0.01$.

the groups in terms of perceived difficulty during self-assessment, $t(232) = 0.83$, $p = .408$, $d = 0.11$, or mental effort during self-assessment, $t(232) = 1.14$, $p = .255$, $d = 0.15$.

RQ 2: What are the effects of performance level descriptions on task performance and instructional efficiency in a revision phase?

Before we addressed our hypotheses regarding performance in the revision phase, we tested whether learners had actually improved their performance in the revision phase. Overall, we found that learners realized substantial performance gains, $F(1, 235) = 109.30$, $p < .001$, $\eta_p^2 = 0.32$. Also, the performance gains on the dimensions the learners chose for revision substantially differed from zero, $t(232) = 9.92$, $p < .001$, $d = 0.65$.

Next, we addressed our hypotheses. In Hypothesis 2a, we assumed that providing learners with rubrics *before* they engage in revision would lead to higher performance gains in a revision phase (Hypothesis 2a). However, a 2 × 2 factorial between-subjects ANOVA with the factors *rubrics vs. rating scales during self-assessment* and *rubrics vs. rating scales during revision* did not reveal a statistically main effect of the factor of *rubrics vs. rating scales during self-assessment*, $F(1, 232) = 0.04$, $p = .835$, $\eta_p^2 < 0.01$. Also, the factor of rubrics vs. rating scales during self-assessment did not affect the level of cognitive load during taking regulation decisions, neither in terms of perceived difficulty, $F(1, 234) = 0.04$, $p = .849$, $\eta_p^2 < 0.01$, nor in terms of invested mental effort, $F(1, 230) = 0.88$, $p = .349$, $\eta_p^2 < 0.01$.

In Hypothesis 2b, we predicted that providing learners with rubrics during revision would not foster performance gains in a revision phase in comparison to providing learners with rating scales. There was no statistically significant main effect of the factor *rubrics vs. rating scales during revision* regarding performance gains in the revision phase, $F(1, 232) = 0.04$, $p = .842$, $\eta_p^2 < 0.01$. Furthermore, there was no statistically significant interaction between the two factors, $F < 1$. To more thoroughly test whether rubrics and rating scales resulted in similar performance gains in the revision phase, in the next step we conducted an equivalence test. Based on recommendations by Lakens (2017), we defined an effect size of $d = 0.10$ as the upper boundary and an effect size of $d = -0.10$ as the lower boundary for effect sizes of interest. To test for equivalence, we determined the 90% confidence interval of the effect size of the difference between the groups with rubric and with rating scales concerning performance gains, which was −0.06 to 0.36. As the interval included a value beyond our pre-set positive boundary, it can be concluded that the performance gains were not equivalent. Together with the aforementioned finding of the ANOVA, there is hence no significant difference but also no equivalence between rubrics and rating scales concerning performance gains in the revision phase.

Similar to our findings regarding performance gains, no significant effects were observed when instructional efficiency, i.e., the performance gains in relation to the experienced cognitive load in the revision phase, was targeted (Hypothesis 2c). Again, but this time contrary to our hypothesis, a 2 × 2 ANOVA did not indicate a statistically significant main effect of the factor of *rubrics vs. rating scales during revision*, $F(1, 227) = 0.20$ $p = .653$, $\eta_p^2 < 0.01$. This pattern of results did not change when, for explorative purposes, only cognitive load in the revision phase and hence not the performance gains in relation to cognitive load (i.e., instructional efficiency) were considered. There were no statistically significant differences between the groups in terms of perceived difficulty during revision, $t(232) = 0.19$, $p = .848$, $d < 0.01$, and mental effort during revision, $t(228) = 0.95$, $p = .343$, $d = 0.13$.

We also analyzed whether the effects of the two factors on performance gains in the revision phase would depend on learners' performance in the initial writing phase. This moderation analysis revealed that the effect of the factor *rubrics vs. rating scales during self-assessment* depended on performance in the initial writing phase, $F(1, 229) = 10.01$, $p = .002$, $\eta_p^2 = 0.04$. Exploration of this moderation effect with the SPSS macro PROCESS 4.0 using the Johnson-Neyman technique revealed that for learners with initial task performance of lower than 18.64, which applied to five participants (i.e., ca. 2% of the sample), rubrics during self-assessment entailed a detrimental effect on performance gains in the revision phase in comparison to rating scales. By contrast, for learners whose performance in the initial writing phase exceeded 32.40, which applied to 31 learners (i.e., 13% of the sample) rubrics during self-assessment beneficially affected performance gains in the revision phase.

For exploratory purposes, we also analyzed whether the findings would be similar when only the performance gains on the dimensions the learners chose to revise were counted. Like for the ANOVA on overall performance gains in the revision phase, there were no significant main effects or interactions between the experimental factors, all $F < 1$. However, again the effect of *rubrics vs. rating scales during self-assessment* depended on learner performance in the initial writing phase, $F(1, 226) = 15.54$, $p < .001$, $\eta_p^2 = 0.06$. Exploration of this moderation effect revealed that for learners with initial task performance of lower than 27.03, which applied to 77 participants (i.e., ca. 32% of the sample), rubrics during self-assessment entailed a detrimental effect on performance gains on the chosen dimensions in the revision phase. By contrast, for learners whose performance in the initial writing phase exceeded 31.70, which applied to 59 learners (i.e., 25% of the sample), rubrics during self-assessment beneficially affected performance gains in the revision phase.

Discussion

Performance level description effects on self-assessment accuracy

In line with our hypotheses, performance level descriptions entailed added value regarding self-assessment accuracy (at least for absolute accuracy). Specifically, in line with Hypothesis 1a, we found that the learners who received full rubrics during self-assessment showed self-assessments of higher absolute accuracy than their counterparts with only rating scales. This is in line with assumptions from the cue utilization framework (Koriat, 1997). Rubrics that include performance level descriptions help learners not only use the given cues (that is, the assessment criteria presented in the rubric) but also to make the best use of it by enabling learners to relate their performance to the benchmarks that are included in the performance level descriptions.

Our findings regarding bias, however, appear to contradict this explanation. Evidently, rubrics did not significantly reduce bias compared to rating scales. One explanation for this lack of effect, however, could be specific limitations of the bias measure. Specifically, the measure of bias has the limitation that positive (i.e., overconfident) and negative (i.e., underconfident) self-assessments can cancel each other out, both within participants (in the case of multiple self-assessments) and across participants. This, in turn, renders interpretation of bias scores challenging and can explain why dissociations between effects on abso-

lute accuracy and on bias can occur. As the beneficial effects of rubrics on self-assessment accuracy were found on the more robust absolute accuracy measure, we hence conclude that performance level descriptions beneficially affect the effects of rubrics on self-assessment accuracy.

Following on from that, and in support of Hypothesis 1b, we found that the quality of learner-generated relations between the degree to which the assessment criteria were met and the performance levels served as a mediator of the benefits in terms of absolute accuracy. Jointly, these results clearly indicate that the component of performance level descriptions is crucial for the benefits of rubrics on self-assessment accuracy to occur. These results suggest that, in terms of self-assessment, rubrics support learners better than the mere exposure of assessment criteria. Specifically, performance level descriptions seem to be essential to help learners inferring their level of performance from their monitoring of the assessment criteria.

Despite these overall conclusive findings, it is important to highlight that the mediating function of the quality of established relations could not explain the whole effect of rubrics on absolute accuracy. There was still a direct effect of condition on self-assessment accuracy, which, on a descriptive level, was even larger than the indirect effect via the quality of established relations. Hence, other factors than the quality of learner-generated relations between the criteria and performance levels contributed to the benefits of performance level descriptions as well. This keeps in line with previous research, which suggests that the processes during self-assessment (and hence, the processes that lead to self-assessment accuracy) are multifaceted (e.g., Andrade et al., 2009; Baars et al., 2014). At this point, however, we could only speculate about these further processes and mechanisms.

Performance level descriptions effects on performance

We did not find strong evidence that performance level descriptions entail substantial added value when it comes to revising one's task performance. The ANOVA showed no statistically significant main effect of the factor *rubrics vs. rating scales during revision* regarding performance gains in the revision phase, hence, not suggesting a positive effect of rubrics here (Hypothesis 2b). Additionally, the equivalence test indicated that the groups were also not equivalent in terms of performance gains in the revision phase. Hence, we cannot clearly state that the conditions are fully equivalent, nor can we see a significantly positive effect – if at all, this effect would be very small. In sum, the results point into the direction that performance level descriptions do not substantially add value in terms of revising a task performance.

Importantly, this lack of effect was not due to the fact that the learners were not able to improve their performance during revision. Overall, we found that learners significantly improved the quality of their written abstracts in the revision phase. However, our analyses indicate that these improvements in performance occurred irrespective of whether learners were supported by full rubrics or only by rating scales during revising their abstracts. One explanation for this finding is that the learners followed a simple "the more assessment criteria I meet, the better"-heuristic during revision and did not orient themselves to specific performance levels in revising their abstracts. These results suggest that providing learners with rubrics during revision may not be adding substantial benefits compared to providing

learners, for example, only with assessment criteria. Hence, in terms of revision behavior, the cues provided by the assessment criteria may be sufficient.

Contrary to our assumptions, however, rubrics during revision were not harmful in that they decreased instructional efficiency (Hypothesis 2c). One explanation for this finding could be that the learners did not only hardly orient themselves to the performance level descriptions in revising their abstracts but scarcely attended to the detailed performance level descriptions at all and hence were scarcely affected by the concomitant (extraneous) processing demands, which might reflected in the lack of effect of rubrics on cognitive load in the revision phase. Future studies that use eye-tracking or think-aloud methodology in order to illuminate learners' activities during revision with rubrics or rating scales could test this tentative explanation.

In terms of our assumption that providing learners with rubrics *before* revision would positively affect revision performance (Hypothesis 2a), our findings were mixed. Specifically, we found that for learners who already performed relatively well in the initial writing phase, receiving rubrics rather than rating scales during self-assessment fostered performance gains during revision. By contrast, for learners who performed relatively poorly in the initial writing phase, the effect reversed and rating scales rather than rubrics during self-assessment were more beneficial. One explanation for this pattern of results could be that the rubrics and rating scales differed in their effects on motivation for learners with low or high performance in the initial writing phase. For learners who already performed well, receiving rubrics that helped them to accurately assign themselves to performance levels may have been motivating, as these learners saw that they already reached relatively high performance levels. Consequently, their self-efficacy was high, and they were motivated to engage in effective revision subsequently. For learners who performed relatively poorly in the initial writing phase, by contrast, the performance level descriptions likely demotivated them, because they resulted in the learners seeing that they had reached only rather low levels of performance. This, in turn, might have decreased their self-efficacy and motivation to engage in effective revision afterwards. Importantly, this pattern is consistent with prior rubric research indicating that learners may respond differently to rubrics depending on their prior performance and academic self-esteem, including evidence that rubrics can lower self-efficacy by making the complexity of high-quality performance salient (Panadero & Jonsson, 2020). Studies in other contexts (e.g., Cleary & Zimmerman, 2001; Trautner & Schwinger, 2020; Zimmerman & Kitsantas, 1997) empirically support this notion, but future studies that replicate the present study and assess the outlined motivational constructs are needed to challenge this tentative explanation.

Limitations

The present study has some important limitations. First, we do not know whether the present effects depend on learners' familiarity with the task. In the present study, we used the task of abstract writing because it is relevant for university students, who need to write abstracts in their theses, for instance, but university students nevertheless have limited expertise and experience with this task. In authentic educational settings, learners' familiarity with the task will often be very heterogeneous, which is why it would be crucial to explore whether task familiarity would moderate the effects found in the present study. For example, it could be assumed that the degree to which learners need support by a rubric with performance

level descriptions would decrease as task familiarity increases, because learners would have already received several instances of feedback from their instructors, which could support them as well. It would be interesting to explore whether learners refer to such *external resources* more frequently in forming self-assessments when they are not yet familiar with a task and whether this aspect would moderate the benefits of providing learners with full rubrics in comparison to rating scales.

A second important limitation is that we tested our hypotheses regarding effects on performance only in a setting in which learners were to revise their task performance. However, this setting is only one of two frequently implemented settings in which beneficial effects of rubrics on performance have been reported in previous research (see Panadero & Jonsson, 2020). That is, rubrics can beneficially affect performance not only in a revision phase but also when learners receive the rubrics right from the beginning of working on a task (e.g., Kocakülah, 2010). When learners receive full rubrics including performance level descriptions right from the beginning of a task, they might be more inclined to orient their activities in performing the task to the different performance level descriptions, because it might be considered as beneficial support that reduces cognitive load. After self-assessment and during revising one's task performance, by contrast, learners might be less inclined to orient to the performance level descriptions, because they already assigned themselves to a performance level (in the self-assessment phase) and might consider the support as redundant to a higher degree. The conclusion that the component of performance level descriptions scarcely affects learner performance should hence not easily be generalized to settings in which learners receive rubrics right from the beginning of working on a task.

A third relevant limitation relates to potential carry-over effects. Exposure to a full rubric during self-assessment may have influenced the learners' behavior during revision, even if only rating scales were present. Although the lack of interaction effects between the two factors indicates that the effects of rubrics in the revision phase did not depend on whether learners received rubrics or rating scales during self-assessment, we cannot rule out certain carry-over effects between the two phases. To address this limitation, in future research it could be worthwhile to address the effects of rubrics on revision without manipulating the presence of rubrics in a previous phase.

Conclusion

In summary, the present study entails the following main contributions. The component of performance level description matters for the effects of rubrics. Evidently, performance level descriptions fostered self-assessment accuracy because they helped learners to infer their level of performance from the monitored cues (i.e., the assessment criteria). By contrast, for fostering task performance, at least when learners receive the rubrics only after an initial phase of working on a task, performance level descriptions scarcely matter. When provided during self-assessment (i.e., before revision), however, performance level descriptions matter for task performance. Evidently, at least for learners who performed already relatively well in the initial writing phase, performance level descriptions fostered effective revision. For learners who performed poorly in the initial writing phase, by contrast, performance level descriptions during self-assessment did not foster – and in some analyses even hindered – subsequent revision. This pattern is consistent with prior evidence that learners respond differently to rubrics depending on their prior performance and academic

self-esteem (Panadero & Jonsson, 2020). We emphasize, however, that this moderation was exploratory and that the comparison concerned rubrics versus rating scales, that is, the presence versus absence of performance level descriptions rather than rubrics as a tool; it therefore requires replication with explicit motivational measures before firm instructional recommendations can be derived.

Appendix

Screenshots from learning environment for self-assessment with rubric

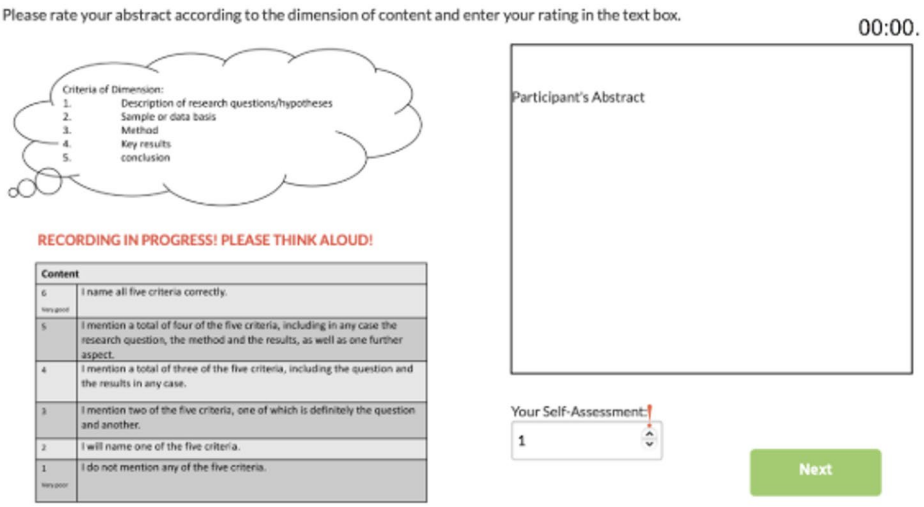
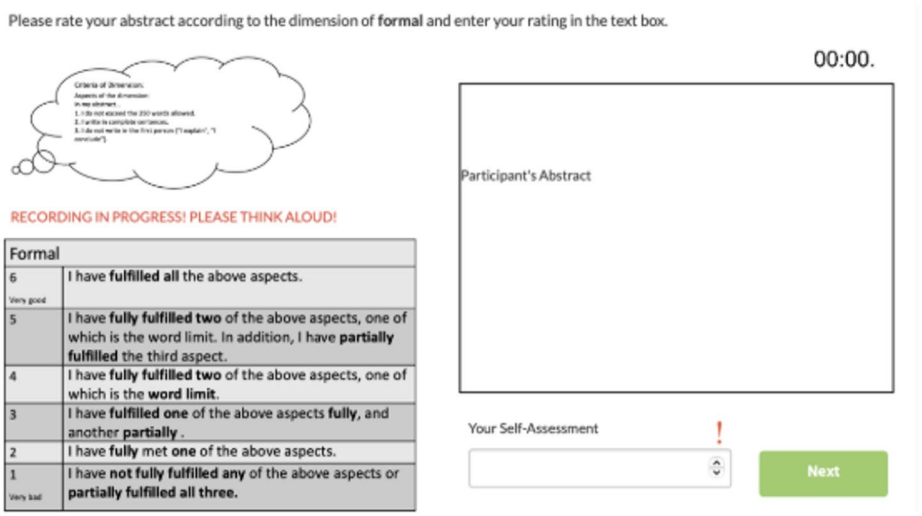
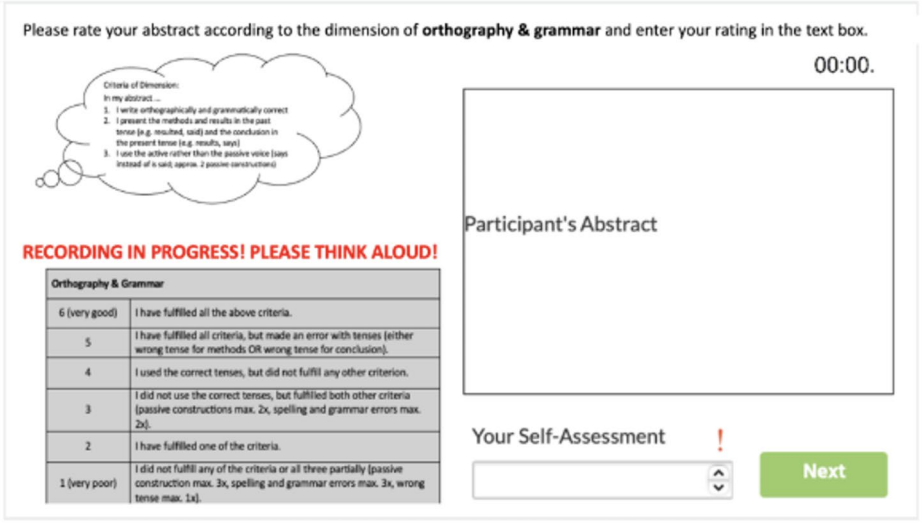


Fig. 5 Self-assessment with rubric for dimension of content



Formal	
6 Very good	I have **fulfilled all** the above aspects.
5	I have **fully fulfilled two** of the above aspects, one of which is the word limit. In addition, I have **partially fulfilled** the third aspect.
4	I have **fully fulfilled two** of the above aspects, one of which is the **word limit**.
3	I have **fulfilled one** of the above aspects **fully**, and another **partially** .
2	I have **fully** met **one** of the above aspects.
1 Very bad	I have **not fully fulfilled any** of the above aspects or **partially fulfilled all three.**

Fig. 6 Self-assessment with rubric for dimension of formal



Orthography & Grammar	
6 (very good)	I have fulfilled all the above criteria.
5	I have fulfilled all criteria, but made an error with tenses (either wrong tense for methods OR wrong tense for conclusion).
4	I used the correct tenses, but did not fulfil any other criterion.
3	I did not use the correct tenses, but fulfilled both other criteria (passive constructions max. 2x, spelling and grammar errors max. 2x).
2	I have fulfilled one of the criteria.
1 (very poor)	I did not fulfil any of the criteria or all three partially (passive construction max. 3x, spelling and grammar errors max. 3x, wrong tense max. 1x).

Fig. 7 Self-assessment with rubric for dimension of orthography & grammar

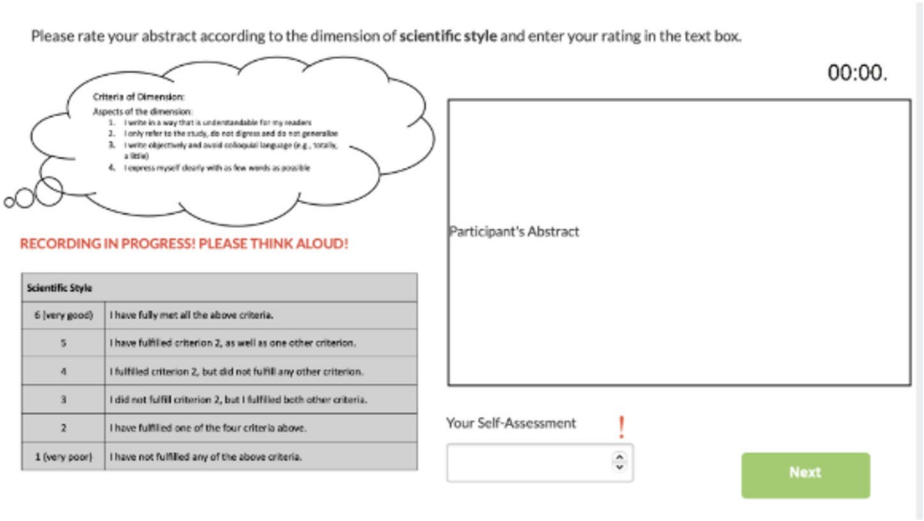


Fig. 8 Self-assessment with rubric for dimension of scientific style

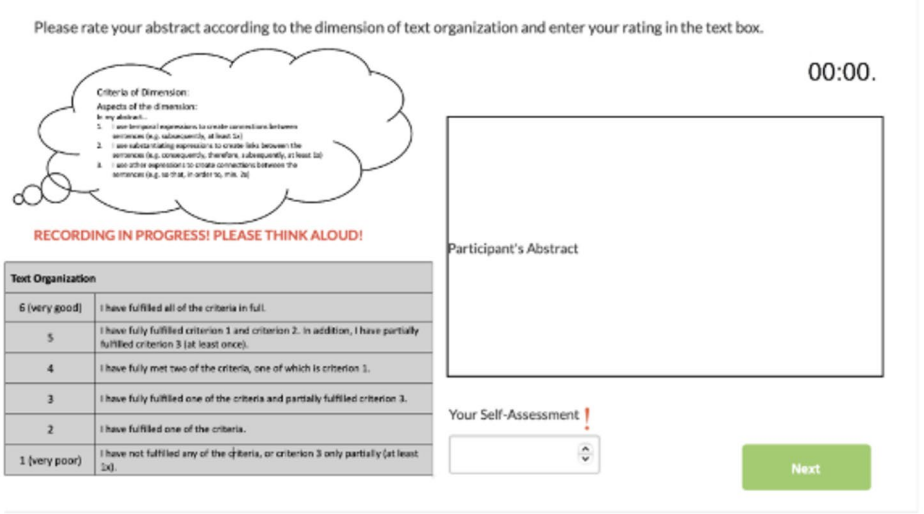


Fig. 9 Self-assessment with rubric for dimension of text organization

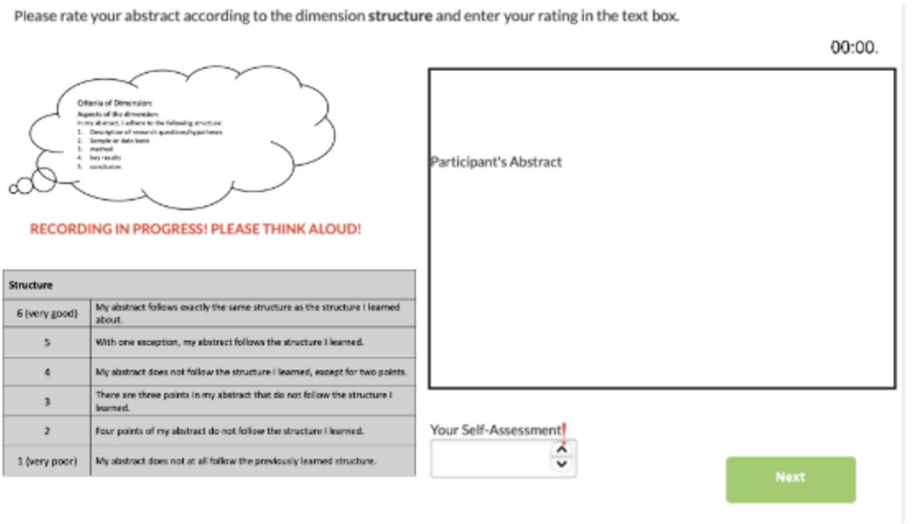


Fig. 10 Self-assessment with rubric for dimension of structure

Appendix B

Screenshots from learning environment for self-assessment with rating scale

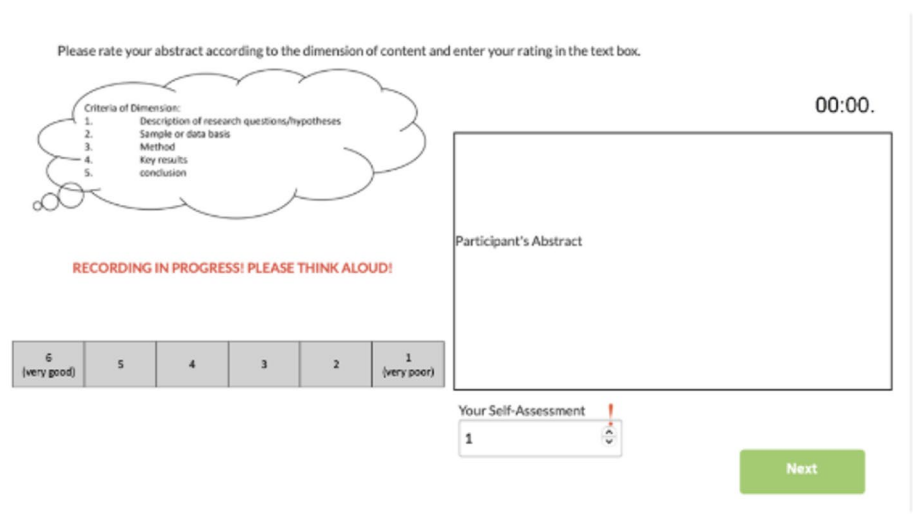


Fig. 11 Self-assessment with rating scale for dimension of content

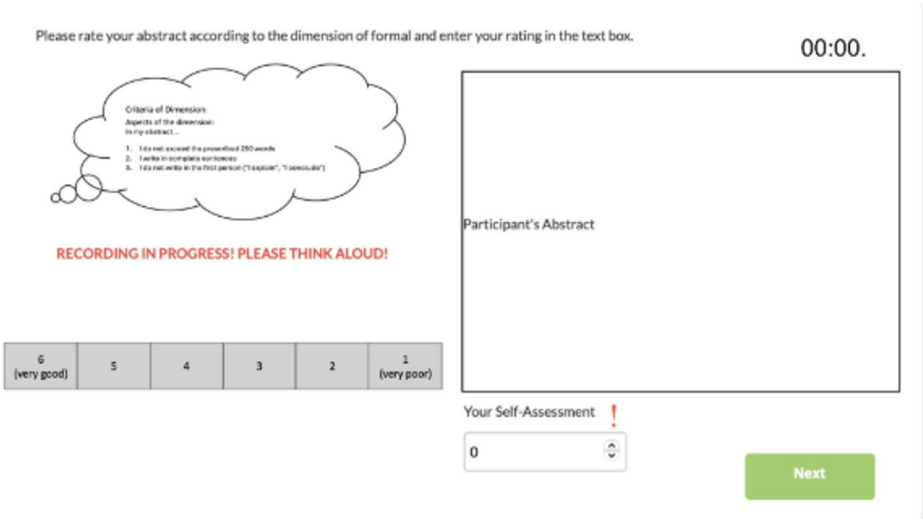


Fig. 12 Self-assessment with rating scale for dimension of formal

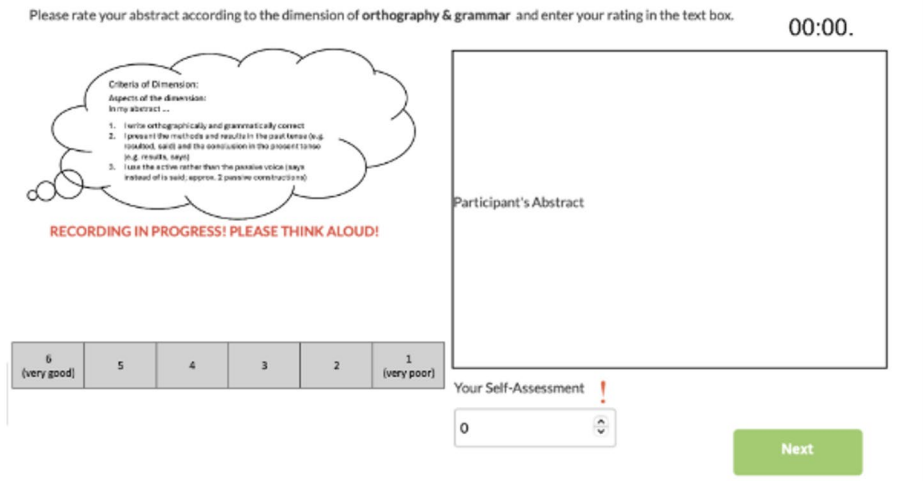


Fig. 13 Self-assessment with rating scale for dimension of orthography & grammar

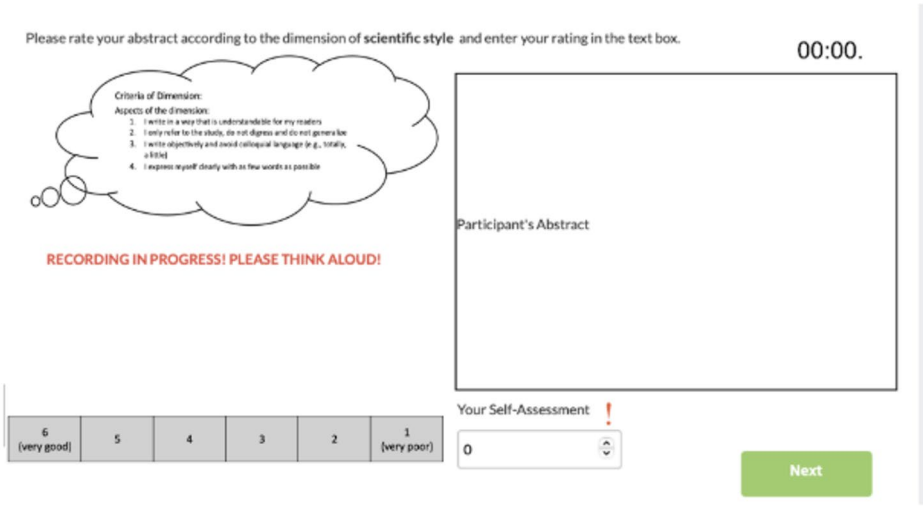


Fig. 14 Self-assessment with rating scale for dimension of scientific style

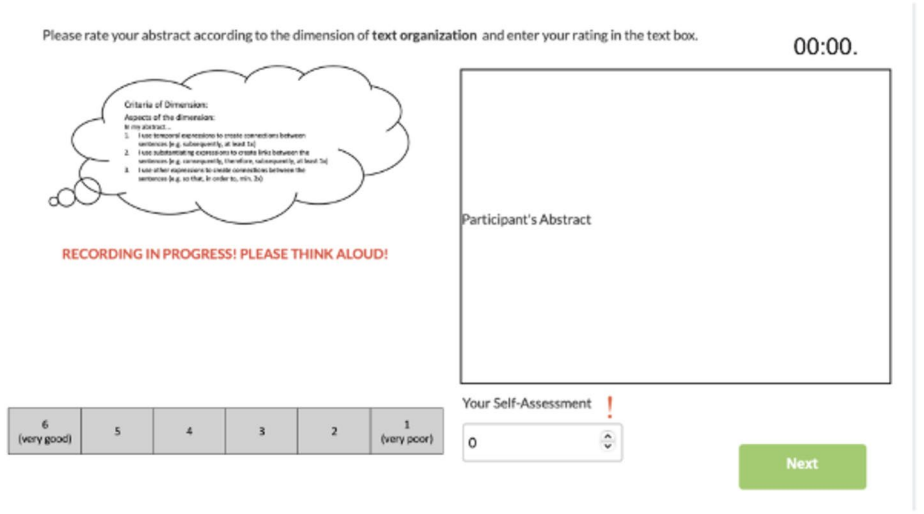


Fig. 15 Self-assessment with rating scale for dimension of text organization

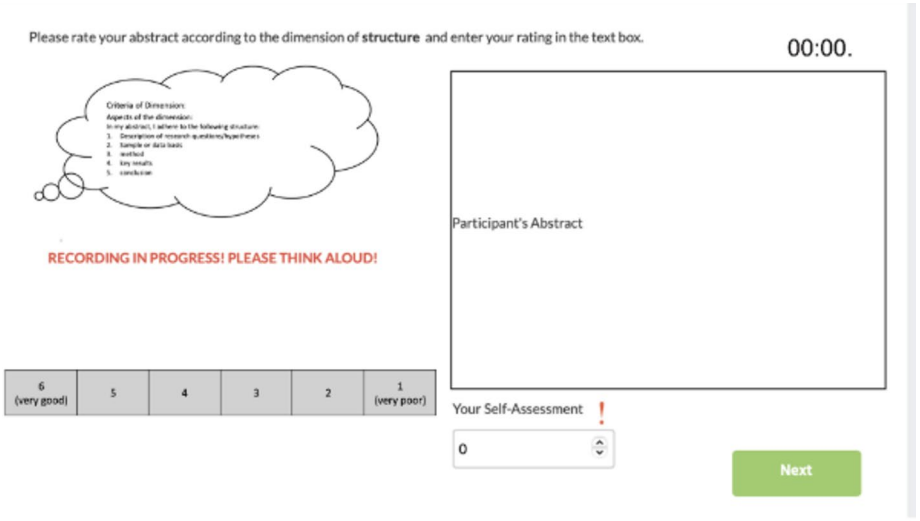


Fig. 16 Self-assessment with rating scale for dimension of structure

Author contributions Conceptualization: Rebecca Krebs, Julian Roelle, Julia Waldeyer, Ernesto Panadero, Björn Rothstein; Methodology: Rebecca Krebs, Julian Roelle, Julia Waldeyer, Ernesto Panadero; Formal analysis and investigation: Rebecca Krebs; Writing - original draft preparation: Rebecca Krebs, Julia Waldeyer, Julian Roelle; Writing - review and editing: Julian Roelle, Julia Waldeyer, Ernesto Panadero, Björn Rothstein; Supervision: Julian Roelle.

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 Data can be found under: https://doi.org/10.17605/OSF.IO/7ZGVR.

Code availability Not applicable.

Declarations

Ethics approval 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. The study was approved by the ethics committee of Ruhr University Bochum (EPE-2022-35).

Consent for publication Not applicable.

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

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. L., Wang, X., Du, Y., & Akawi, R. L. (2009). Rubric-referenced self-Assessment and self-efficacy for writing. *The Journal of Educational Research*, *102*(4), 287–302. https://doi.org/10.3200/JOER.102.4.287-302

Baars, M., Vink, S., Van Gog, T., De Bruin, A. B. H., & 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 students' 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

Balan, A., & Jönsson, A. (2018). Increased explicitness of assessment criteria: Effects on student motivation and performance. *Frontiers in Education*, *3*, 81. https://doi.org/10.3389/feduc.2018.00081

Bol, L., Riggs, R., Hacker, D., Dickerson, D., & Nunnery, J. (2010). The calibration accuracy of middle school students in match classes. *Journal of Research in Education*, *21*, 81–96.

Brookhart, S. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, *3*, 1–12. https://doi.org/10.3389/feduc.2018.00022

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review, 67*(3), 343–368. https://doi.org/10.1080/00131911.2014.929565

Busch-Lauer, I. (2012). Abstracts – eine facettenreiche Textsorte der Wissenschaft [Abstracts – a multifaceted type of scientific text]. *Linguistik Online*, *52*(2), 6–22. https://doi.org/10.13092/lo.52.293

Cleary, T. J., & Zimmerman, B. J. (2001). Self-Regulation differences during athletic practice by experts, non-experts, and novices. *Journal of Applied Sport Psychology*, *13*(2), 185–206. https://doi.org/10.1080/10413200175314988 3

De Bruin, A. B., Dunlosky, J., & Cavalcanti, R. B. (2017). Monitoring and regulation of learning in medical education: The need for predictive cues. *Medical Education*, *51*(6), 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 judgements 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

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. MIT Press.

Fastré, G. M. J., Van der Klink, M. R., & Van Merriënboer, J. J. (2010). The effects of performance-based assessment criteria on student performance and self-assessment skills. *Advances in Health Sciences Education*, *15*(4), 517–532. https://doi.org/10.1007/s10459-009-9215-x

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

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

Hattie, J. A. C. (2009). *Visible Learning. A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *13*(4), 351–371. https://doi.org/10.1002/(SICI)1099-0720)13:4%3C351::AID-ACP589%3E3.0.CO;2-6.

Kocakülah, M. S. (2010). Development and application of a rubric for evaluating students' performance on Newton's laws of motion. *Journal of Science Education and Technology*, *19*, 146–164. https://doi.org/10.1007/s10956-009-9188-9

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

Krebs, R., Rothstein, B., & Roelle, J. (2022). Rubrics enhance accuracy and reduce cognitive load in self-assessment. *Metacognition and Learning, 17*(2), 627–650. https://doi.org/10.1007/s11409-022-09302-1

Krebs, R., Waldeyer, J., Rothstein, B., & Roelle, J. (2024). Do rubrics enhance self-assessment accuracy? *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie*, *56*(3), 144–158. https://doi.org/10.1026/0049-8637/a000296

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lerdpornkulrat, T., Poondej, C., Koul, R., Khiawrod, G., & Prasertsirikul, 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/10.1177/0734282917747423

Paas, F. (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. G., & Van Merriënboer, J. J. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors*, *35*(4), 737–743. https://doi.org/10.1177/001872089303500412

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1–4.https://doi.org/10.1207/S15326985EP3801_1

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

Panadero, E., Alonso-Tapia, J., & 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., Pinedo, L., & Fernández-Castilla, B. (2023). Effects of rubrics on academic performance, self-regulated learning, and self-efficacy: A meta-analytic review. *Educational Psychology Review*, *35*(4), 113. https://doi.org/10.1007/s10648-023-09823-4

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

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

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

Swales, J. M., & Feak, C. B. (2009). *Abstracts and the writing of abstracts* (Vol. Vol. 2). University of Michigan Press ELT. https://doi.org/10.3998/mpub.309332

Swales, J. M., & Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press. The Cambridge Applied Linguistics Series.

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.

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

Trautner, M., & Schwinger, M. (2020). Integrating the concepts self-efficacy and motivation regulation: How do self-efficacy beliefs for motivation regulation influence self-regulatory success? *Learning and Individual Differences, 80*, Article 101890. https://doi.org/10.1016/j.lindif.2020.101890

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., Endres, T., Roelle, J., Baars, M., & Renkl, A. (2024). How to optimize self-assessment accuracy in cognitive skill acquisition when learning from worked examples. *Educational Psychology Review*. https://doi.org/10.1007/s10648-024-09944-4

Zimmerman, B. J., & Kitsantas, A. (1997). Developmental phases in self-regulation: Shifting from process goals to outcome goals. *Journal of Educational Psychology*, *89*(1), 29.

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