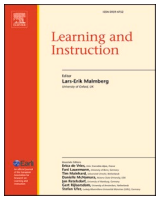
Learning and Instruction 99 (2025) 102168

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc



Putting excellence first: How rubric performance level order and feedback type influence students' reading patterns and task performance

Ernesto Panadero [a,b,c], Pablo Delgado [d], David Zamorano [b,*], Leire Pinedo [b], Alazne Fernández-Ortube [b], Lucía Barrenetxea-Mínguez [b]

[a] Centre for Assessment Research, Policy and Practice (CARPE), Institute of Education, Dublin City University, St. Patrick's Campus, Dublin, Ireland
[b] Education Regulated Learning and Assessment (ERLA group), Facultad de Educación y Deportes, Universidad de Deusto, Bilbao, Spain
[c] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[d] Laboratorio de Diversidad, Cognición y Lenguaje, Universidad de Sevilla, Spain

ARTICLE INFO

Keywords:
Rubric
Feedback
Reading patterns
Eye-tracking
Academic performance

ABSTRACT

Background: Rubrics are structured assessment tools that describe criteria and levels of performance, helping students understand expectations and improve their work. They are widely used to support learning in educational settings. However, little is known about how students process rubrics in real time, and empirical research on rubric design and feedback effects is limited.
Aim: This study examines how university students engage with rubrics during two landscape analysis tasks, focusing on two variables: the order of performance levels (highest first vs. last) and the type of feedback received (no feedback [control], process-based, product-based, or rubric-based). By combining eye-tracking and think-aloud protocols, the study offers a multimodal perspective on students' visual attention and cognitive engagement.
Sample: Eighty undergraduate students from six degree programs were randomly assigned to one of four feedback conditions.
Methods: A randomized controlled trial was conducted. Eye-tracking data—fixation times, number of visits, and gaze transitions—and verbal data from think-aloud protocols were collected across task phases. Integrating these process-tracing methods enabled detailed analysis of how students interacted with the rubric and how engagement related to performance.
Results: Students focused primarily on the highest performance level, especially when it appeared first. Visual attention to this level predicted task performance; verbal references did not. Rubric-based feedback increased visual alignment between rubric and task, while process-based feedback led to the strongest performance gains.
Conclusion: Rubric design and feedback type significantly influence student engagement and performance. Eye-tracking and think-aloud data provide complementary insights, reinforcing rubrics' instructional value when paired with targeted feedback.

Rubrics are structured assessment tools that articulate evaluation criteria across multiple levels of performance (Brookhart, 2018). Widely used in educational settings, they are designed to support both instruction and learning and have been shown to enhance student performance (Panadero et al., 2023). Their effectiveness is often attributed to the increased transparency they provide, enabling students to understand expectations, self-assess, and improve their learning (Brookhart & Chen, 2015; Panadero and Jonsson, 2013). Importantly, rubrics may facilitate student learning by increasing the transparency of performance expectations, which can help learners focus their attention more effectively during task completion. Some studies have suggested that this clarity may also reduce cognitive load by supporting more efficient information processing (Krebs et al., 2022). Additionally, rubrics can be more efficient than other assessment tools such as exemplars (Lipnevich et al., 2022). Thus, they are generally beneficial, although recent evidence also points to conditions under which their effects may be limited or mixed (Panadero et al., 2023). To maximize the benefits of rubrics, it is crucial to understand when and how they exert their

* Corresponding author. ERLA office, Universidad de Deusto, Bilbao, 48007. Spain.
E-mail address: david.zamorano@deusto.es (D. Zamorano).

https://doi.org/10.1016/j.learninstruc.2025.102168
Received 16 August 2024; Received in revised form 21 May 2025; Accepted 23 May 2025
Available online 30 May 2025

positive effects, so they can be designed and implemented in the most effective way.

Despite their popularity (Dawson, 2017; Panadero et al., 2023), little is known about how students engage with rubrics while performing academic task (Panadero and Jonsson, 2020). In particular, there is a lack of research examining how students read rubric components, allocate attention across performance levels, and incorporate rubric information into their task strategies. Further, while feedback is often assumed to support rubric use (Brookhart, 2018), there is limited empirical evidence on how different feedback types interact with rubric engagement and performance (Panadero and Jonsson, 2020). These gaps call for research that examines the cognitive and attentional processes involved in rubric use, as well as how design features and feedback conditions shape those processes.

1. Rubrics and their effects

A rubric can be defined as a tool that: "articulates expectations for student work by listing criteria for the work and performance level descriptions across a continuum of quality" (Brookhart, 2018). Rubrics are typically represented as tables or matrices, with the first column detailing assessment criteria and the subsequent columns indicating varying performance levels, which may range from high to low quality or vice versa (Panadero et al., 2023).

Rubrics can be effective for both summative and formative purposes (e.g., Brookhart, 2018; Jonsson & Svingby, 2007); however, their use by students tends to align more with formative intentions. Panadero and Jonsson (2013), in their systematic review, argued that rubrics significantly enhance transparency by making criteria and standards explicit. They further posited that rubrics contribute to improved academic performance, better reflection on feedback, enhanced planning and revision of assignments, and reduced anxiety and negative self-regulation. Subsequent reviews, such as Brookhart and Chen (2015), and a recent meta-analysis (Panadero et al., 2023), have supported these findings.

Importantly, simply providing students with a rubric may not always result in improved performance (Wollenschläger et al., 2016), especially if they lack prior experience with rubrics or the specific task at hand (Panadero et al., 2023). To maximize their effectiveness, rubrics should be instructionally integrated into the learning process (Andrade, 2005). Thus, achieving impactful student use of rubrics requires not only careful design but also thoughtful pedagogical support. In this study, we focus on two such support features: how rubrics are designed (i.e., the performance level order) and how they are pedagogically accompanied (i.e., through different types of feedback). Importantly, while not all feedback conditions in this study include explicit references to the rubric in the feedback, all students had equal access to the rubric throughout the assigned task.

2. Rubric design: the order of performance level

One of the key elements that could influence the effects of rubrics is how they are designed (Brookhart, 2018). While they are basic instruments usually represented as a table, depending on the content and organization of their element they could be easier to read and use (Jonsson & Svingby, 2007). For instance, if the language used in a rubric is too technical, it may hinder student comprehension. Similarly, the number of performance levels must be carefully chosen: too many can make the rubric difficult to navigate, while too few may fail to capture the full range of possible performances. As Brookhart (2018) argues, the number of levels should be aligned with the specific decisions required and the reliable distinctions that can be identified and utilized in students' work. Similarly, Jonsson and Svingby (2007) note that fewer levels often increase reliability in evaluations, though they might oversimplify performance distinctions. Unfortunately, these two articles are reviews and have not directly investigated the ideal number of performance levels. To our knowledge, there is scant research exploring how the design of rubrics influences their use, particularly from the students' perspective, as the only study we know of focused on the use of rubrics by teachers (Humphry & Heldsinger, 2014).

A critical aspect of rubric design that could significantly affect how students interact with and understand them is the order of presentation of the performance levels. Particularly, the placement of the highest performance level, typically the most detailed, may influence how students approach the rubric. Questions arise such as whether students predominantly focus on the highest level or if their engagement with the rubric varies depending on its position. To investigate this, we will manipulate the order of presentation from highest to lowest and vice versa.

3. Feedback as an implementation factor

Given that rubrics are instructional scaffolds, it is to be expected that their effects are influenced by instructional variables. One that seems crucial is feedback (Andrade, 2005; Wollenschläger et al., 2016). Here, an interesting dilemma arises as rubrics themselves are tools to provide feedback (Lipnevich et al., 2022). This raises an important distinction: while rubrics inherently provide evaluative information, students may still require additional feedback to learn how to interpret and act upon that information effectively. Essentially, students -particularly those not yet skilled in self-assessment- rely on external feedback to maximize the benefits of rubric use. Without this essential feedback, there is a significant risk of students misapplying rubrics, which can impede their ability to fully realize and develop their capabilities.

Previous research has explored how giving different types of feedback and the use of a rubric influence variables such as academic performance, self-regulatory strategies, or metacognition (Panadero et al., 2012; Saiz-Manzanares et al., 2017; Wollenschläger et al., 2016). For example, Panadero and colleagues (2012) compared the effects of two types of feedback, mastery vs performance, when students were using rubrics in a landscape analysis. They found that mastery feedback, increased their participants self-efficacy, whereas performance feedback worsened the students' negative self-regulatory strategies. However, these studies did not specifically investigate how feedback influences the students' interaction with the rubric itself. Thus, further research in this area is necessary.

In this study, we compared four feedback conditions: no feedback (control), process-based feedback (focused on task execution strategies), product-based feedback (focused on the final output), and rubric-based feedback (aligned with performance levels in the rubric).

4. Students' rubric use: tracking visual attention and cognitive strategies

Understanding how students use rubric information, both in terms of how they visually attend to performance criteria and how they interpret and apply this information cognitively, requires methodological tools that go beyond outcome measures. This is particularly important when considering how design features (e.g., performance level order) and instructional supports (e.g., feedback type) may shape not only how rubrics are used, but also the cognitive strategies and attentional patterns that underlie their use. To investigate these internal processes, it is necessary to adopt a methodological approach capable of capturing students' real-time use of rubrics as they perform academic tasks.

This need for process-oriented investigation is underscored by prior research on rubric use. While some studies have explored how different raters or teachers engage with rubrics (e.g., Postmes et al., 2023; Winke & Lim, 2015), less attention has been paid to how students engage with and make sense of these tools in real time during task execution (e.g., Andrade & Du, 2007; Turley & Gallagher, 2008).

Most existing research has focused on retrospective self-reports about rubric use (e.g., Andrade & Du, 2005), which offer limited

insight into the actual processes involved. In contrast, process-tracing methods such as eye-tracking and think-aloud protocols provide complementary access to students' real-time cognitive and attentional engagement. Eye-tracking captures students' visual attention patterns, revealing which rubric components they fixate on, in what sequence, and for how long (Holmqvist et al., 2011). However, visual behavior alone does not reflect the full range of cognitive processing. For this reason, in this study we also integrate think-aloud protocols (Ericsson & Simon, 1993), which encourage students to verbalize their thoughts and strategies as they work with the rubric, providing insights into how they interpret and apply rubric information during the task.

Despite their potential, these methods have rarely been used to study how students engage with rubrics. To our knowledge, only one published study has used eye-tracking in a rubric context (Winke & Lim, 2015), focusing on teacher-raters rather than students and without counterbalancing rubric structure. Moreover, no study to date has combined both eye-tracking and think-aloud protocols to examine how students read and use rubrics. This gap underscores the need for empirical work that explores both visual attention and cognitive strategies in parallel, especially when students are asked to use rubrics formatively during complex academic tasks.

5. A multimodal process-tracing approach to rubric use

To address this methodological gap, our study adopts a multimodal approach that triangulates visual and verbal process data (Panadero, 2023). Eye-tracking provides fine-grained, real-time information on attentional focus (Jarodzka et al., 2021), which we analyze using three indicators: fixation times, frequency of visits to specific rubric elements, and gaze transitions between rubric and task stimuli. These metrics allow us to infer how students engage with rubric content during key moments of task performance (van Gog & Scheiter, 2010).

Complementing this, we employ think-aloud protocols to examine students' explicit reasoning and cognitive strategies. These verbal data capture dimensions of understanding and rubric use that may not be evident from visual attention patterns alone, such as how students compare performance levels, interpret rubric criteria, or plan their task responses (Ericsson & Simon, 1993).

The integration of these two data sources follows a logic of triangulation common in mixed-methods research (Creswell & Plano-Clark, 2018), allowing us to capture both convergences and divergences in students' engagement with rubrics (Panadero, 2023). This methodological strategy enhances the interpretive validity of our findings and provides a richer, more nuanced understanding of how rubrics function as instructional tools during authentic learning activities.

6. The present study: aim, research questions, and hypotheses

This study investigates how university students engage with rubrics during two landscape analysis tasks, focusing on two key instructional variables: the order in which performance levels are presented in the rubric (highest first vs. lowest first) and the type of feedback received (no feedback -control-, process-based, product-based, or rubric-based). Our aim is to examine how these two instructional variables students engagement with rubrics (i.e., reading patterns and use) and their task performance. To capture both attentional and cognitive engagement, we combine two complementary process-tracing methods: eye-tracking and think-aloud protocols. Eye-tracking provides measures of visual behavior, focusing here on fixation times, number of visits, and gaze transitions, while think-aloud protocols offer insight into students' verbalized reasoning as they use the rubric. Specifically, we address four research questions (RQ) and hypotheses (H).

RQ1. How do students engage with performance levels (PL) of the rubric during their initial interaction?

H1 We hypothesize that students will prioritize the highest performance level (PL4) in both their visual and verbal engagement with the rubric.

RQ2. Does the order of PL in the rubric affect students' reading patterns?

H2 We hypothesize that presenting the highest performance level (PL4) first will more strongly guide students' attention toward it compared to when it appears last.

RQ3. Is students' rubric reading pattern related to their performance on a first task?

H3 We hypothesize that students' attention to PL4, especially visual engagement, will be positively associated with task performance.

RQ4. How does the type of feedback on students' first task performance influence students' engagement with the rubric and performance in a subsequent task?

H4 We hypothesize that rubric-based feedback will enhance students' engagement with the rubric, particularly toward PL4, and lead to greater improvement in task performance compared to other feedback types or no feedback.

7. Method

7.1. Participants

7.1.1. Inclusion and exclusion criteria

We reached to students from six different undergraduate programs from the same university. The only exclusion criterion was having visual impairments or a recent eye surgery. One participant was excluded on this basis, and all remaining participants had normal or corrected-to-normal vision.

7.1.2. Participants

Our participants were 138 first year university students. They were randomly assigned to one of four experimental conditions using a random number generator. his approach aimed to balance participant characteristics across conditions.

Eye-tracking data were collected while participants engaged in the experimental tasks. Due to the strict quality controls applied to ensure data validity, 58 participants (42 %) were excluded from the analyses. These exclusions did not substantially affect the group composition, as the remaining participants were still relatively evenly distributed across the four experimental conditions (20, 19, 21, and 20 participants, respectively). The final sample thus consisted of 80 participants. Of these, audio recordings from four participants were not successfully captured due to technical issues, making their think-aloud data unavailable. Nevertheless, their eye-tracking data were valid and included in the analyses.

The participants in the final sample analyzed here were majoring in Psychology (43.75 %), Medicine (18.75 %), Education (12.5 %), Sport Sciences (12.5 %), Social Work (8.75 %), and Social Education (3.75 %). The average achievement level of the sample before entering university was 8.3 over 10, reflecting a notably high academic standard among the participants.

Finally, considering the rubric and the experiment was conducted in Spanish we explored the dominant language of the participants. They self-reported their dominant language as can be seen at Table 1, with an even distribution of the participants.

Table 1
Distribution of values for self-reported dominant language, mother tongue, and language in compulsory education.

Language status variable	Values	*n* in 1–4 PL order condition	*n* in 4-1 PL order condition
Self-reported dominant language	Spanish	20	23
	Euskera-Spanish	19	18

7.1.3. Sampling procedure

Participants were recruited through convenience sampling from undergraduate programs at the institution where several of the authors (first, third, fourth, fifth, and sixth) are affiliated. We approached the students through emailing their teachers as to attend one of the sessions to present our study. There, the aim and significance of the study was explained, and all the students were invited to take part. A sheet of paper was held to the students so the ones willing to participate could write their email address. In total, we approached seven classroom groups, from which 174 students allowed us to contact them and 138 came to the laboratory to participate. As already mentioned, due to data loss derived from filtering eye-tracking data quality (see *Apparatus and measures* section below), only 80 (58 %) were taken as participants in this publication, 55 of them females. Participants that fully completed the experiment received an economic incentive (5€) for taking part in the study.

8. Materials and intervention

8.1. Rubric

The rubric was created in a previous study (Panadero et al., 2012), using Social Science experts' models of landscape analyses (see Appendix A for more information on the design and implementation). The rubric contains four performance levels and five assessment criteria (see Appendix B). This allowed us to examine whether the position of performance levels influenced how students engaged with the rubric. To control for potential confounding effects, we also counterbalanced the vertical order of assessment criteria across participants. The rubric was written in Spanish. In Appendix C provides additional procedural details regarding the feedback process, including the rubric-based feedback.

We explored the structure and reliability of the rubric via three measures. First, we examined the underlying factor structure of the rubric by conducting an exploratory factor analysis (EFA) using the Unweighted Least Squares extraction method and Oblimin rotation. Prior to conducting the EFA, the suitability of the data for factor analysis was evaluated. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was 0.738, indicating moderate adequacy for factor analysis. Bartlett's test of sphericity was significant ($x^2(10) = 65.73$, $p < .001$), confirming that the correlation matrix was suitable for factor analysis. The analysis revealed one factor with an eigenvalue greater than 1, which explained 40.90 % of the variance. Following extraction, this factor accounted for 26.51 % of the variance, suggesting a unidimensional structure. Overall, the results suggest that the rubric can be represented by a single underlying construct. The moderate communalities and factor loadings indicate that the criteria are sufficiently related to justify their inclusion in a single-factor model, although the variance explained is modest.

Secondly, we measured the internal reliability of the five assessment criteria comprising the rubric using the omega coefficient (McDonald, 1999), as it is considered more appropriate than Cronbach's alpha for analyzing items with fewer than five possible values (Trizano-Hermosilla & Alvarado, 2016). The results showed omega values of 0.603 and 0.631 for the grading of students' written analysis 1 and written analysis 2, respectively. Although modest, these values are within an acceptable range for educational instruments with a small number of items and performance levels, particularly in formative assessment contexts (Lance et al., 2006; Loewenthal & Lewis, 2020).

Finally, we employed the rubric to evaluate students' task performance, with two independent raters (the fifth and sixth authors) achieving excellent inter-rater reliability (see the section below, "Task Performance").

8.1.1. Task: landscape analysis

A landscape analysis typically involves a systematic examination and evaluation of the physical, environmental, and socio-economic aspects of a particular area or region. Therefore, the task consisted in analyzing, first orally and then in writing, two landscapes with different characteristics: a) a rural area with Continental climate and (b) a rural area with Mediterranean climate (see Appendix D). Each landscape was presented twice: (1) along with the rubric while the participant performed the oral analysis and (2) then along with the rubric and a text processor (i.e., word file) where participants wrote their final landscape analysis. Importantly, the ability to analyze landscapes is a competence included in the secondary education curriculum in Spain.

8.1.2. Instructions

All participants received the same oral instructions for the task (Appendix E), and were also given a printed sheet summarizing the task requirements to ensure consistency across sessions.

8.1.3. Types of feedback

We assigned participants to four feedback conditions using stratified randomization. In the no feedback condition, participants received no information after completing the two landscape analyses. In the process feedback condition, participants received comments about how they approached the task (i.e., analyzing the landscape), focusing on process-level feedback (Hattie & Timperley, 2007) that compared their approach to that of an expert and highlighted key omissions (e.g., "This aspect appears in the expert analysis, but you didn't mention it …"). In the product feedback condition, participants received task-level feedback (Hattie & Timperley, 2007) on the quality of their written output (i.e., the landscape analysis), without referencing the rubric (e.g., "You correctly addressed all required elements in this section."). Finally, in the rubric-based feedback condition, comments were explicitly anchored in the rubric, linking performance to specific criteria and levels (e.g., "For this criterion, you would be at level X because you commented on …"). At no point did we provide feedback on how to use the rubric itself (e.g., "You should focus more on the performance level 4"). All feedback was provided immediately after task completion to ensure timeliness and relevance while minimizing recall issues. See Appendix C for a more detailed explanation and excerpts of the different types of feedback.

8.2. Apparatus and measures

8.2.1. Eye tracker

All the stimuli were presented on a 24-inch screen with 1680x1050-pixel resolution. Participants' eye movements were recorded using a Tobi Pro Fusion screen-based eye tracker at 250 Hz that was calibrated using a five-point scheme (minimum required accuracy: 0.5°; Holmqvist et al., 2011). Calibration was performed for each participant before starting each experimental phase (see Fig. 2). See Appendix F for detailed information on the eye-tracker settings and data cleaning.

We measured participants' visits to each rubric PL (i.e., areas of interest) and total fixation time on each rubric cell in three testing times: (1) initial reading of the rubric, (2) oral analysis 1, and (3) oral analysis 2 (see Fig. 2). In addition, we measured participants' number of gaze transitions between each rubric PL and the picture of the landscape (see Fig. 1) during both oral analysis tasks (see Fig. 2).

Visits to each rubric PL were identified by tracking gaze transitions from any part of the stimulus (either the rubric or the picture) to a

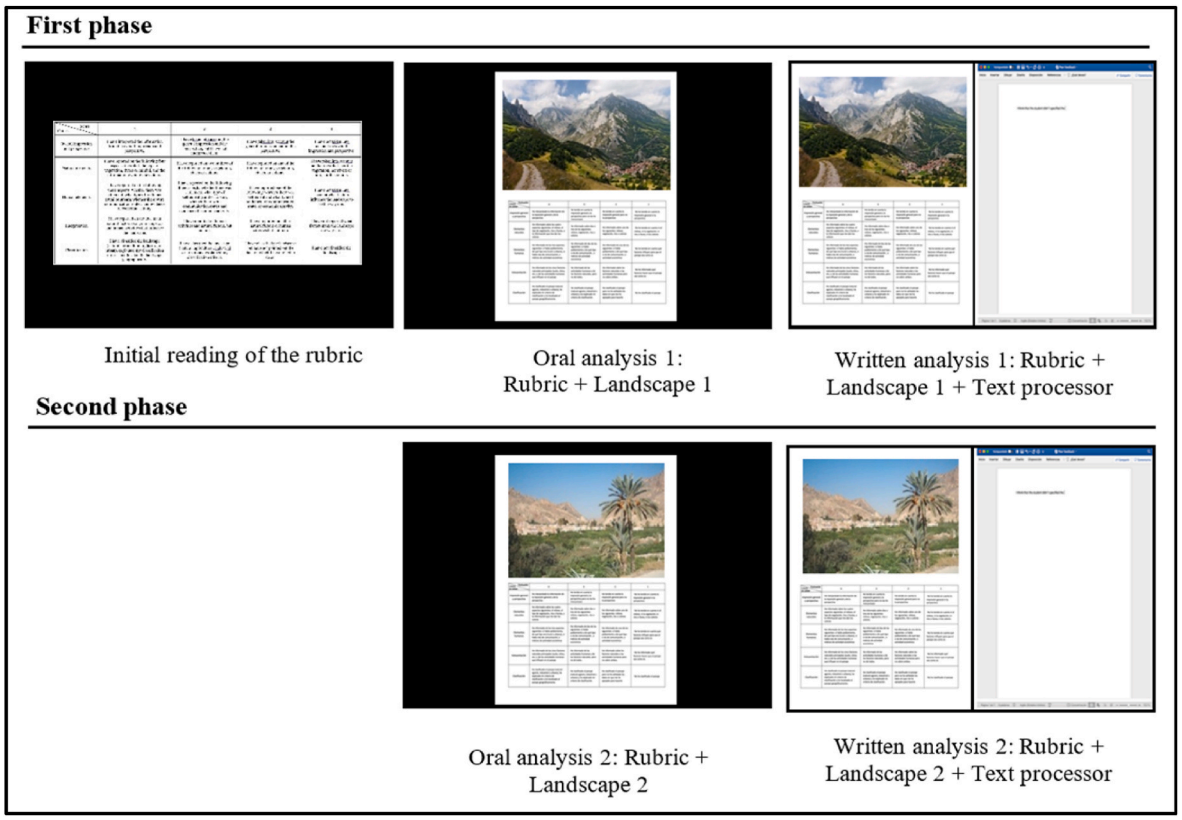


Fig. 1. Visualization of the process and materials used by participants.

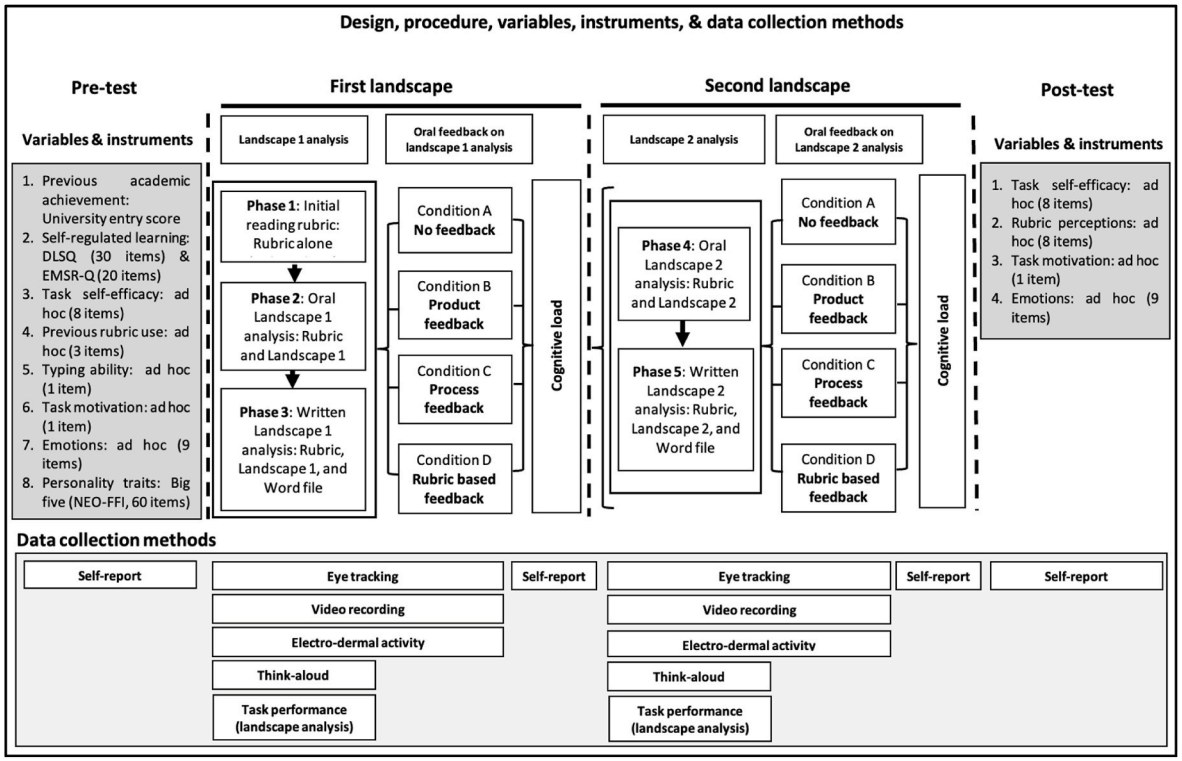


Fig. 2. A detailed illustration of the procedure.

specific PL. Single-fixation visits to a PL were excluded to avoid spurious fixations during transitions between areas of the stimuli (e.g., if a participant read PL4 and then PL1 but made a fixation on PL2 in between, that fixation was excluded).

Regarding fixation times on each cell, they were divided by the number of characters in each rubric cell to account for differences in text

length across cells. While characters are not inherently meaningful units, word length is a well-established determinant of fixation duration (Salthouse & Ellis, 1980). Controlling for character count thus allowed us to better isolate the effects of rubric content and structure on students' visual attention, as recommended in text complexity and eye-tracking research (Feng et al., 2010).

Gaze transitions between each rubric PL and the picture were identified by tracking eye movements in both directions: from the picture to a rubric PL and vice versa.

8.2.2. Think-aloud method

To examine students' cognitive engagement with rubrics, we collected concurrent think-aloud protocols (TAP) as participants analyzed the landscapes (see Procedure for more details). Below, we outline the key methodological decisions taken regarding the implementation, processing, and analysis of these data.

Think-Aloud Instructions and Prompting. Before starting the experiment, participants received a brief training session on how to think aloud. They were asked to complete a familiar visualization exercise: "Imagine you are standing at the front door of your house. Now, as you walk inside, count the number of windows your house has while verbalizing everything you see, do, and think." This exercise was intended to familiarize participants with verbalizing cognitive processes without engaging in additional interpretation or justification (Ericsson & Simon, 1993).

Throughout the experimental phases, participants were reminded to verbalize their thoughts as they interacted with the rubric and task materials. If they remained silent for 30 s, the researcher provided a standardized prompt: "What are you doing now?" or "Remember to tell me what you are thinking, feeling, and the strategies you are using." This 30-s rule aligns with established think-aloud procedures (McDonald and Petrie, 2013) to maintain verbalization without disrupting cognitive flow. Importantly, during the written landscape analysis, prompting was minimized to avoid interfering with cognitive processes. In those cases, prompts were only delivered if the participant appeared disengaged or had potentially finished the task.

Cognitive Load and Task Design. The landscape analysis task was chosen to ensure a moderate level of cognitive demand. This aligns with prior recommendations suggesting that tasks that are too simple may only elicit superficial cognitive processing, while overly complex tasks can overwhelm participants and interfere with their ability to verbalize their thoughts (Charters, 2003; Ericsson & Simon, 1993). Additionally, the task was aligned with the social sciences training students had received in secondary education. This ensured that they were familiar with the type of analytical reasoning required, while still being challenged to engage in higher-order cognitive processing.

Levels of Verbalization. We aimed to elicit verbalizations corresponding to Level 1 and Level 2 of Ericsson and Simon's (1993) classification, as these are the least intrusive and most natural forms of verbalization during task performance. Level 1 verbalizations occurred when participants naturally described their visual and cognitive processes in real-time, such as stating which rubric section they were reading. Level 2 verbalizations included brief elaborations, such as explaining how they were using rubric information to analyze the landscape. Since this study did not focus on deep metacognitive reflections, Level 3 verbalizations, which require additional interpretation or justification beyond the natural thought process, were not elicited as part of the protocol.

Triangulation with Eye-Tracking Data. This study integrates TAP with eye-tracking measures—specifically fixation times, number of visits, and gaze transitions—to examine students' engagement with rubric elements. Although both data sources were collected concurrently, our primary analytical strategy treated them as complementary but independent indicators: visual behavior and verbalized reasoning were analyzed in parallel to identify converging or diverging patterns across groups and conditions. In some cases, we conducted cross-checks to ensure that verbalizations referring to specific rubric elements were preceded by visual attention to those areas; however, these checks were used selectively to validate protocol quality rather than as a systematic basis for analysis. Thus, the triangulation in this study reflects convergence at the level of interpretive patterns, rather than fine-grained, synchronized episode-level coding.

Segmentation and Unit of Analysis. For analysis, we segmented think-aloud verbalizations into discrete units based on expressed ideas rather than rigid sentence boundaries. Each unit was determined by shifts in focus, such as moving from describing a rubric section to discussing its relevance to the task. This flexible segmentation allowed for a more precise mapping of verbalized strategies to eye-tracking data.

Coding Process and Inter-Rater Reliability. Think-aloud data were coded based on three analytical categories. First, direct references to rubric elements (e.g., mentioning a specific performance level). Second, comparisons between performance levels (e.g., statements contrasting two levels). Third, indications of strategic reading behavior (e.g., statements suggesting sequential or selective reading). The coding process followed an iterative approach: the first and fourth authors independently coded 52 think-aloud segments, reaching a Cohen's kappa of 0.80, indicating strong inter-rater reliability (Landis & Koch, 1977). The remaining segments were coded collaboratively, with both authors working side by side and resolving the only two discrepancies through consensus.

8.2.3. Task performance

Task performance was evaluated independently for each of the two written landscape analyses. The same rubric used throughout the study was applied for scoring (see Appendix B). The fifth and sixth authors independently scored two rounds comprising 15 % of the written tasks to establish inter-rater reliability, which reached an excellent level (ICC = 0.99; Intra-class Correlation Coefficient, Hallgren, 2012). Once reliability was established, the sixth author scored the remaining analyses. Any discrepancies between coders were discussed and resolved collaboratively.

8.2.4. Students' University Entry Exam scores

The students reported their University Entry Exam scores. In the Spanish education system, these scores vary between 7 (minimum score needed to access Bachelor programs) and 14 (maximum potential score).

8.2.5. Additional measures not analyzed in the present study

In addition to the instruments and measures reported above, the participants self-reported other variables (e.g., self-regulated learning) by means of questionnaires. These data are not part of the scope of the present study, so they are not analyzed here. The description of these questionnaires can be found in Appendix G.

8.3. Experimental design

We employed an experimental mixed design including one within-participant factor and two between-participant factors. The order of the performance levels (PLs) in the rubric was manipulated as a between-participant factor with two conditions: (1) highest to lowest (PL4–PL1) and (2) lowest to highest (PL1–PL4). Although all participants were exposed to the same four performance levels (PL1 to PL4), we examined their visual and verbal engagement with each PL, which were treated as repeated measures in the statistical analyses. Additionally, participants were randomly assigned to one of the four feedback conditions: (1) control group without feedback, (2) process feedback, (3) product feedback, and (4) feedback based in the rubric. Participants were also assigned to the experimental conditions ensuring a balanced gender and previous academic performance representation. We distributed the participants evenly among the four experimental conditions.

8.4. Procedure

Participants were recruited through an in-class presentation delivered by one of the research team members. During this presentation, students were informed about the general objectives of the study and invited to voluntarily express their interest in participating by writing their email address on a sign-up sheet. Only students who signed up were contacted to schedule a date and time for their individual participation. Upon arrival at the laboratory, participants were again given a detailed explanation of the procedure, including the nature of the tasks and the data to be collected (e.g., eye-tracking, think-aloud protocols). At that point, they were invited to read an informed consent document and were explicitly told they could decline or withdraw from the study at any point without consequence. Only those who agreed signed the consent form and proceeded with the experiment.

Each participant attended the experimental session individually and was welcomed by a member of the research team. The study was conducted entirely in Spanish, including all materials and instructions. After providing informed consent and receiving a detailed briefing about the procedure, participants were trained on how to think aloud and completed a brief practice task. They then filled out a set of initial questionnaires. Once completed, they were seated—separated by a room divider to ensure privacy—facing the computer equipped with a built-in eye tracker. A five-point calibration was performed for each participant. The experimental procedure consisted of two phases, as illustrated in Fig. 2.

First, participants were shown the rubric and instructed to read it carefully and familiarize themselves with its structure and content -this constituted the first phase. No further guidance was provided, as the objective was to observe how students would interact with the rubric in the absence of explicit instructions. Importantly, all participants had prior experience working with rubrics. While reading the rubric, participants were reminded to think aloud and verbalize their thoughts, feelings, and strategies.

Once they finished reading the rubric, participants were shown the same rubric, but this time along with the first landscape (rural area with Continental climate) and they were asked to analyze the landscape aloud -this constituted the second phase. Lastly, a text processor was opened, and participants were asked to write the landscape analysis they already developed aloud -this constituted the third phase. While performing the written landscape analysis, they could consult the rubric and landscape as they were also presented along with the text processor.

After finishing the written landscape analysis, feedback was orally provided immediately by either the third, fourth or fifth authors. The content of the feedback depended on the condition to which the participant was randomly assigned: no feedback, to the process, to the product, or rubric based. The corresponding feedback was delivered from a document that contained detailed feedback as to ensure same content within conditions. The researcher working with the participant delivered the pieces of information that were relevant based on the participant's performance so that the feedback was tailored to the participant. After this, the eye tracker recording stopped as the first landscape analysis had finished. Then the participant filled out a cognitive load scale, whose data are not presented here as they fall outside the scope of the current study, which focuses exclusively on students' visual and verbal engagement with rubrics.

After completing the cognitive load scale, participants proceeded directly to the second landscape analysis. The eye-tracker was recalibrated, and the same procedure was repeated with two key modifications: (1) the rubric was not shown in isolation again, as this step was only necessary during the first phase; and (2) a different landscape was used—this time, a rural area with a Mediterranean climate. Participants first completed an oral analysis of this second landscape -this constituted the fourth phase-, followed by a written analysis -this constituted the fifth phase. After completing the task, they received oral feedback from one of the authors, depending on their assigned condition. Finally, participants filled out a new set of questionnaires.

8.5. Data analyses

We performed linear mixed-effect models (LMM) to analyze students' fixation times and number of visits to each PL, whereas students' rubric-picture transitions and performance in the written analysis tasks were analyzed by means of generalized linear models (GLM) using R software version 4.3.0 in all cases. Eye movements during the initial rubric reading (phase 1) and during oral analysis 1 (phase 2; see Fig. 2) were combined to examine a more comprehensive picture of participants' rubric reading and use when using the rubric during the first task. Variables with non-normal distribution (i.e., skewness or kurtosis values do not fall within the ±2 range; George & Mallery, 2010) were $\log_{10}$-transformed.

Decisions on setting fixed or random slopes of the random effects for each LMM were based on null models' goodness of fit comparisons. An overview of the models and a more detailed description of these analyses and the R functions used can be found in Appendix H.

For the analysis of think-aloud protocols, we focused on those segments in which participants explicitly referenced performance levels (PLs) or engaged in rubric-related reading behavior. These protocols were manually coded and categorized based on their content, distinguishing between direct references to specific PLs, sequential reading strategies, and general or structural engagement with the rubric. Frequency counts for each category were computed, and distributions were analyzed using Chi-Square tests or one-way ANOVA, depending on the nature of the comparison and the structure of the data. In cases where participants contributed multiple protocols, independence of observations could not always be assumed; this limitation was acknowledged in the interpretation of results. When relevant, additional analyses excluding high-frequency contributors were conducted to assess the robustness of the findings.

9. Results

The results presented below offer an integrated summary of the key findings from eye-tracking and think-aloud data in relation to each research question. For transparency and completeness, the full statistical outputs and detailed model information are available in Appendix I.

RQ1. How do students engage with performance levels (PL) of the rubric during their initial interaction?

We first examined how frequently students visited each performance level (PL) in the rubric and how much time they spent reading them, using eye-tracking data from phases 1 and 2. Linear mixed-effect model (LMM) analyses showed that rubric order had no significant effect on the number of visits to all PLs combined ($t = -0.28$, $p = .09$) nor on the total fixation time ($t = -1.80$, $p = .08$). In contrast, clear differences emerged based on cell PL. Participants visited PL4 significantly more often than PL1, PL2, and PL3 (t values > 6.08, p values < 0.001). Regarding fixation time, students spent significantly more time on PL4 compared to PL2 and PL3 ($t = 2.76$, $p = .01$; and $t = 4.10$, $p < .001$, respectively), while time on PL4 and PL1 was not significantly different ($t = 1.13$, $p = .28$).

Next, we analyzed the frequency with which students verbally referenced specific performance levels in their think-aloud protocols during Phase 1. Of the 174 protocols reflecting reading actions, 76 included explicit references to a particular PL. Initial analysis showed a relatively even distribution across levels: 19.7 % (PL1), 23.7 % (PL2), 27.6 % (PL3), and 28.9 % (PL4), with no significant differences detected ($\chi^2(3) = 1.58$, $p = .66$). However, when excluding three participants who produced a disproportionately high number of references, the adjusted dataset ($n = 26$) showed that PL4 was referenced significantly more than the other levels ($\chi^2(3) = 11.85$, $p < .01$).

Taken together, eye-tracking and think-aloud data converge in

indicating a heightened level of student attention to PL4 during initial rubric engagement. While fixation time for PL4 did not significantly differ from PL1, PL4 was visited more frequently and, after outlier adjustment, referred to more often in verbal protocols. This triangulated evidence suggests that PL4 served as a key anchor point in students' rubric processing.

RQ2. Does the order of performance levels in the rubric affect students' reading patterns?

First, we analyzed whether the order in which performance levels (PL) were presented in the rubric influenced students' reading patterns, using both the number of visits and fixation times from phases 1 and 2. Linear mixed models revealed significant interaction effects between rubric PL order and the specific PL visited. Participants who received the 1–4 PL order visited PL1 and PL2 significantly more often than those who received the 4–1 order ($t = 2.66$, $p = .01$; and $t = 2.35$, $p = .02$, respectively); conversely, participants in the 4–1 condition visited PL4 more frequently than those in the 1–4 condition ($t = 3.82$, $p < .001$; see Fig. 3). Fixation times showed a parallel pattern (see Fig. 4): participants in the 1–4 condition spent significantly more time on PL1 and PL2 ($t = 4.45$, $p < .001$; and $t = 2.80$, $p = .01$, respectively); conversely, participants in the 4–1 condition spent significantly more time on PL4 than their peers in the 1–4 condition ($t = 2.23$, $p = .03$). Additionally, participants in the 1–4 condition focused more on PL1, with significant differences for PL1 versus PL3 and PL4 ($t = 4.06$, $p < .001$; and $t = 3.19$, $p = .003$, respectively); while participants in the 4–1 condition spent more time fixating on PL4 than on the other levels (vs PL1: $t = 3.52$, $p < .01$; vs PL2: $t = 3.69$, $p < .001$; vs PL3: $t = 2.69$, $p = .01$).

Next, we examined whether rubric PL order influenced students' verbalized references to performance levels in their think-aloud protocols. A Chi-Square Test of Independence on all 76 references did not yield significant results, $\chi^2(3) = 5.72$, $p = .13$. A follow-up analysis excluding three participants with disproportionately high numbers of mentions ($n = 26$) also showed no significant effect, $\chi^2(3) = 1.23$, $p = .75$. These results suggest that rubric order did not affect how frequently students verbally referenced the performance levels.

Taken together, eye-tracking data indicate that students' visual attention was strongly influenced by the order of performance levels in the rubric: they spent more time and revisited more frequently the performance level that was positioned first. However, this pattern was not mirrored in their verbalizations, which did not show significant variation across rubric order conditions. While visual behavior appears sensitive to structural presentation, students' verbal references to performance levels were comparatively stable. Therefore, our hypothesis is only partially supported: although students who received the 4–1 rubric order focused more on PL4 and relied less on lower levels -as expected- their verbalized strategies did not reflect this shift in attention.

RQ3. Is students' rubric reading pattern related to their performance on a first task?

First, we explored whether students' rubric reading behavior was associated with their task performance in the written landscape analysis. General linear models (GLM) revealed that among all performance levels, only attention to PL4 predicted higher performance. Specifically, the number of visits to PL4 significantly predicted better task outcomes ($t = 2.11$, $p = .04$), as did the average fixation time on PL4 ($t = 2.56$, $p = .01$). No such associations were observed for PL1, PL2, or PL3. Rubric order showed no significant effect on performance, and neither the interaction between rubric order and number of visits nor the interaction between rubric order and fixation time reached significance. Finally, although the number of gaze transitions between PL4 and the picture during oral analysis 1 did not reach conventional significance, it showed a positive trend ($t = 1.75$, $p = .08$), suggesting that shifting attention between PL4 and the task may also be functionally related to performance.

Then, we examined whether students' verbal references to specific performance levels in think-aloud protocols (TAP) were linked to task performance. A total of 85 protocols from Phases 1 and 2 included explicit references to PLs and were included in the analysis. A one-way ANOVA revealed no significant differences in task performance as a function of the referenced PL, $F_{(3, 81)} = 0.60$, $p = .62$, with homogeneity of variances confirmed by Levene's test. A follow-up analysis excluding

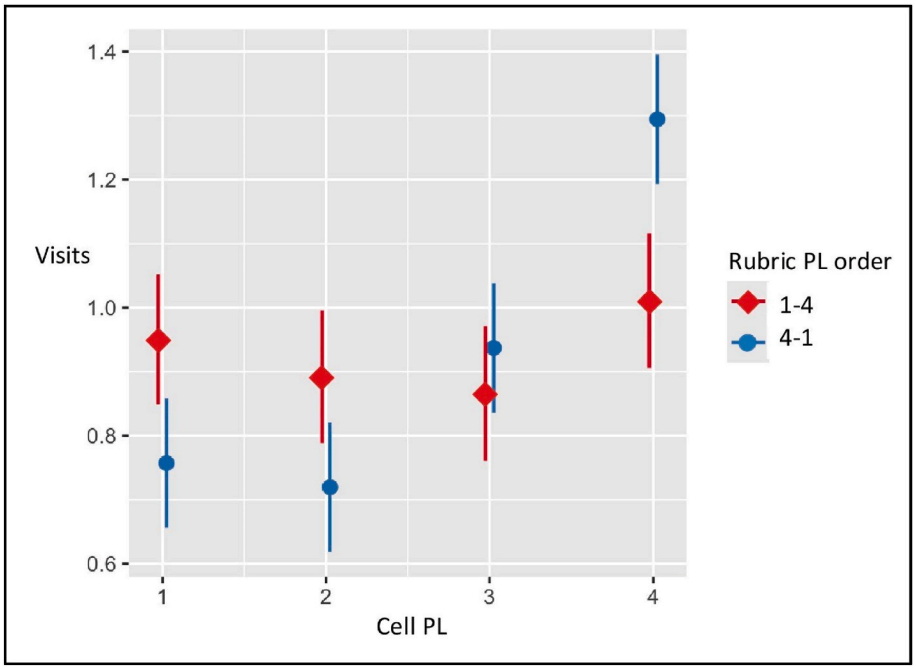


Fig. 3. Students' frequency of visits to each PL per rubric PL order condition.
Note. In the figure, cell PL is orderer from 1 to 4 in this figure for both rubric PL orders, but note that the order was inverted (i.e., from 4 to 1) for the students who read the rubric with 4-1 PL order (values in red) when performing the experimental task. Also note that estimates are based on log_{10}-transformed data, they thus do not represent number of visits.

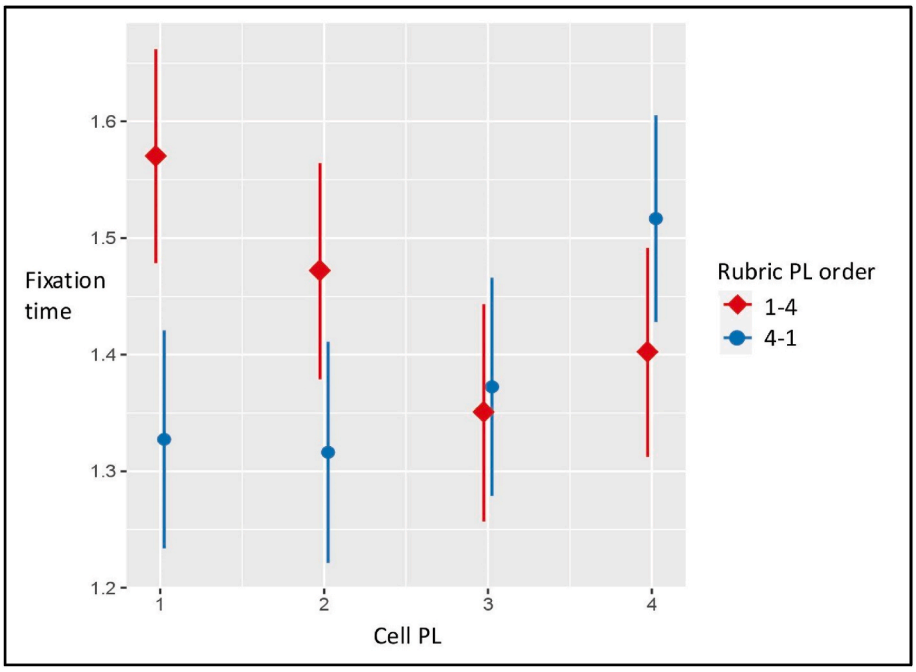


Fig. 4. Students' fixation times for each cell PL and rubric PL order condition.
Note. In the figure, cell PL is orderer from 1 to 4 in this figure for both rubric PL orders, but note that the order was inverted (i.e., from 4 to 1) for the students who read the rubric with 4-1 PL order (values in red) when performing the experimental task. Also note that estimates are based on log_{10}-transformed data, they thus do not represent milliseconds.

three participants who accounted for a disproportionate number of TAPs yielded similar results, $F_{(3,\ 30)} = 1.05$, $p = .39$. Post hoc comparisons showed no significant pairwise differences between groups. It is important to note that this analysis does not fully meet the assumption of independence, as some participants contributed multiple TAPs.

Taken together, the results partially support our hypothesis. Eye-tracking data showed that both the frequency and duration of visual attention to PL4 were positively associated with better performance in the written task, suggesting that engaging with the rubric's depiction of high-level performance plays a functional role in guiding students' output. In contrast, verbal references to performance levels did not predict task performance, indicating a potential dissociation between what students look at and what they verbalize. Overall, the findings suggest that attention to PL4, as evidenced through gaze behavior, may serve as a meaningful marker of productive rubric use, while verbal protocols appear less diagnostic in this context.

RQ4. How does the type of feedback on students' first task performance influence students' engagement with the rubric and performance in a subsequent task?

First, we examined whether the type of feedback received after the first written analysis influenced students' rubric reading behavior during the second oral analysis (Phase 4). We focused on PL4, given that previous analyses identified it as the performance level most strongly associated with task performance. General linear models (GLMs) revealed that neither feedback type nor its interaction with rubric PL order had a significant effect on the number of visits to PL4 (all $ts < 0.43$, $ps > 0.67$) or on fixation time (all $ts < 0.55$, $ps > 0.58$), suggesting similar levels of visual engagement with PL4 across feedback conditions. However, when analyzing changes in PL4-picture gaze transitions (i.e., the difference between oral analyses 1 and 2), a significant effect of feedback was found. Specifically, only students who received rubric-based feedback significantly increased their transitions from PL4 to the picture ($t = 2.19$, $p = .03$), whereas no such increase was observed in the other feedback conditions. This effect was not moderated by rubric PL order. Furthermore, among all gaze-related measures at Phase 4, only the number of PL4-picture transitions significantly predicted students' performance in written analysis 2 ($t = 2.01$, $p = .049$), even after controlling for University Entry scores.

Next, we analyzed verbal engagement with the rubric at Phase 4 which was minimal. Out of 1621 think-aloud protocols collected, only eight included explicit references to performance levels. Due to the extremely small and uneven distribution of these verbalizations across feedback conditions, no statistical analysis was conducted for this data.

Finally, we analyzed the extent to which feedback type influenced students' performance in written analysis 2. A score-change index was calculated by subtracting performance on written analysis 1 from that on written analysis 2. The GLM showed a significant main effect of feedback. Students who received process-based or product-based feedback improved significantly more than those in the control group ($t = 3.51$, $p < .001$; and $t = 2.51$, $p = .01$, respectively). The rubric-based feedback group also showed gains, but the difference only approached significance compared to the no-feedback group ($t = 1.98$, $p = .05$). No interaction effect with rubric PL order was observed.

Taken together, these results provide partial support for H4. As predicted, students who received rubric-based feedback showed increased PL4-picture transitions, which also positively predicted task performance, suggesting deeper cognitive engagement with performance criteria. However, rubric-based feedback did not lead to increased visits or fixation times on PL4, nor did it result in significantly higher verbal engagement—likely due to floor effects in the TAP data. Finally, in line with H4c, all three feedback conditions produced gains in performance, but contrary to expectations, rubric-based feedback yielded the smallest (and marginal) improvement compared to process- and product-based feedback.

10. Discussion

This study investigated how two key elements in rubric use—namely, the design of the rubric (specifically, the order in which performance levels are presented) and the type of feedback provided—influence students' engagement with the rubric and their task

performance. To examine these questions, we analyzed four research questions (RQ), drawing on eye-tracking data, think-aloud protocols, and students' task outcomes across a multi-phase experimental design.

Our results provide several converging findings. First, students directed most of their attention, both visually and verbally, to the highest performance level (PL4), particularly during their initial engagement with the rubric (RQ1). Second, the order in which performance levels were presented had a clear effect on reading behavior: participants prioritized the level that appeared first, but this structural manipulation did not translate into differences in verbal engagement or performance outcomes (RQ2). Third, the depth of attention to PL4 -especially gaze transitions between PL4 and the task-was associated with better performance, highlighting the functional role of visual engagement with exemplar-level information (RQ3). Finally, rubric-based feedback led to a specific increase in PL4-task gaze transitions during a subsequent task, while process- and product-based feedback resulted in stronger performance gains overall (RQ4).

Importantly, this study offers novel methodological insight into how students engage with rubrics—an area that remains largely underexplored. By combining two process data methods -eye-tracking and think-aloud protocols-it provides a more nuanced and multimodal account of students' interactions with rubric features. This integration allowed us to identify consistencies (e.g., students' prioritization of PL4 across modalities) as well as discrepancies (e.g., stronger predictive value of visual behavior compared to verbalized strategies), highlighting the value of using multiple data streams to understand the cognitive and behavioral processes involved in rubric use.

10.1. Rubric design

Our findings show that the design of a rubric -specifically, the order in which performance levels are presented-substantially influences how students read and use the rubric. Presenting the highest performance level (PL4) first consistently attracted the greatest amount of attention, both in terms of number of visits and fixation time. Moreover, reading PL4 was positively associated with task performance, while no such relationship was found for the other levels. These results suggest that placing PL4 first may enhance efficiency by directing students' attention to the most informative performance level. This structural feature might reduce unnecessary processing demands, although we did not assess cognitive load directly in this study; this remains a plausible mechanism for future investigation. This alignment likely facilitates more efficient processing, as proposed by models of goal-directed attention (Friedman & Förster, 2001). Interestingly, although rubric design affected how students read the rubric, it did not produce direct effects on task performance. We interpret this as a boundary condition of the task demands: the impact of rubric structure may become more evident in more cognitively complex or less familiar academic tasks.

While there is extensive theoretical literature on rubric design (e.g., Arter & McTighe, 2000; Brookhart, 2013, 2018; Isaacson & Stacy, 2009; Mertler, 2001), empirical studies examining how specific design features influence rubric usability and effectiveness remain limited. Among the few exceptions, Humphry and Heldsinger (2014) demonstrated that rubric structure can create a "halo effect," whereby judgments are skewed not by bias but by the layout itself, and that structural changes can mitigate this. Similarly, Papageorgiou et al. (2015) highlighted the trade-off between classification reliability and meaningful differentiation, recommending an optimal number of six performance levels to balance accuracy and usability. Both studies underscore the need for empirical evidence to guide the development of rubrics with both psychometric and pedagogical integrity.

Our study contributes to this limited empirical base by focusing on how students -rather than teachers or raters-engage with rubric structure during task performance. In contrast to Winke and Lim's (2015) work with expert raters, our design was counterbalanced and focused on authentic student use. Moreover, our approach to adjusting fixation time by character count, rather than word count, provides a more precise metric of visual attention. Overall, our findings support the idea that structural features of rubrics shape user engagement (as predicted in H2), even if those effects do not always translate directly into improved performance.

These findings also align with longstanding recommendations in the rubric literature. Jonsson and Svingby (2007) emphasize the importance of clarity and usability in rubric design, while Brookhart (2018) highlights the need to balance detail with simplicity to minimize cognitive overload. Our results offer empirical support to these theoretical positions, showing how a seemingly minor design element—the order of performance levels—can measurably influence how students engage with rubrics.

10.2. Feedback effects: a key variable in rubric implementation

The type of feedback students received after their first written task influenced both their engagement with the rubric and their subsequent performance. Students who received rubric-based feedback showed a specific increase in gaze transitions between PL4 and the task, suggesting more deliberate cross-referencing of rubric expectations during the second task. This finding is consistent with the idea that rubric-based feedback can sharpen students' attention to performance criteria and support more informed task execution. However, this effect did not generalize to other indicators of engagement: there were no differences in the number of visits to PL4 or time spent fixating on it across feedback conditions. Moreover, rubric-based feedback did not produce the strongest gains in task performance.

Instead, process-based and product-based feedback were the most effective in improving students' scores on the second written task. This aligns with prior findings showing that detailed, process-oriented feedback enhances learning outcomes more reliably than feedback focused solely on standards or evaluative language (Panadero et al., 2012; Wollenschläger et al., 2016). These results suggest that while rubric-based feedback may direct attention toward relevant features of performance, it may lack the scaffolding required for students to transform that awareness into action.

More broadly, the role of feedback in rubric implementation has been widely discussed from theoretical and practical perspectives (e.g., Andrade, 2005; Brookhart, 2018; Jones et al., 2017; Moskal, 2000; Panadero et al., 2012; Turley & Gallagher, 2008). However, there remains a clear need for empirical research that investigates how different types of feedback influence the way students engage with rubrics during actual task performance (Panadero et al., 2023). While studies have explored rubric use by teachers and raters with empirical rigor (e.g., Bresciani et al., 2009; Postmes et al., 2023), fewer have examined how feedback shapes rubric reading behavior among students—a gap this study begins to address.

In sum, our results indicate that feedback type is not only a powerful moderator of student outcomes but also shapes the nature of students' engagement with the rubric itself. These findings contribute to a more nuanced understanding of rubric implementation, underscoring the need for feedback that is not only aligned with the rubric but also actionable from the learner's perspective.

10.3. Triangulation of process data: convergences and divergences

A key contribution of this study lies in the integration of two process-tracing methods: eye-tracking and think-aloud protocols. This triangulation allowed us to capture complementary dimensions of student engagement with the rubric, revealing both consistencies and divergences in how students attended to, interpreted, and verbalized their interaction with performance criteria (e.g., Creswell & Plano-Clark, 2018).

In several cases, eye-tracking and TAP data converged. For instance, both data sources indicated a strong focus on PL4 during initial rubric

reading, supporting the idea that students prioritize the highest performance level when orienting themselves to task expectations. However, in other instances, the two data streams diverged. Eye-tracking measures were more sensitive to experimental manipulations (e.g., rubric order, feedback type), whereas verbalizations did not consistently reflect these differences. Moreover, only gaze-based indicators -particularly transitions between PL4 and the task-predicted performance, suggesting that visual attention may be a more reliable proxy of goal-directed engagement than verbalized reasoning in time-constrained tasks (van Gog & Scheiter, 2010).

These findings reinforce the methodological value of combining multiple process measures in educational research. While verbal protocols provide insights into explicit cognitive strategies (Ericsson & Simon, 1993), eye-tracking captures moment-by-moment attentional patterns that may reflect more automatic or unconscious aspects of learning behavior (Jarodzka et al., 2021). The use of both allowed us not only to validate our interpretations across modalities, but also to highlight when and how the modalities diverge—offering a more comprehensive picture of how students engage with assessment tools such as rubrics.

10.4. Educational implications

This study offers several actionable insights for educators seeking to make rubrics more effective tools for learning. First, placing the highest performance level first may help students engage more efficiently with rubric criteria, especially in time-constrained settings. From an instructional perspective, this finding highlights the importance of aligning rubric design with students' natural information-processing tendencies. By positioning goal-oriented information first, educators can support learners in quickly forming mental representations of what quality looks like, which may help them better plan, monitor, and evaluate their work. This design choice may also be particularly beneficial for students with lower self-regulatory skills or limited prior experience with assessment tools, as it reduces the need to scan the rubric in full to understand expectations, a hypothesis to be tested. In this sense, small adjustments to rubric layout can become low-cost, high-impact strategies to scaffold student autonomy and self-assessment.

Second, while rubrics are widely used as tools for both assessment and feedback, our findings suggest that their effectiveness is significantly enhanced when combined with individualized verbal feedback. Students in the control condition—who received only the rubric without any feedback—showed the weakest performance outcomes. Interestingly, students who received rubric-based feedback did show increased visual alignment between their rubric use and the task, suggesting deeper engagement. However, this engagement did not translate into the highest performance gains. In contrast, the process- and product-based feedback groups outperformed both the control and rubric-based conditions in terms of task performance. These results indicate that while rubrics help structure and clarify expectations, they are not sufficient on their own. For students to benefit fully, rubrics must be embedded within broader instructional strategies that include rich, tailored feedback to support students in bridging the gap between current and desired performance.

Finally, the differentiated effects observed across gaze behavior and verbal reports highlight the importance of teaching students not just what rubrics are but how to use them strategically. Simply providing a rubric does not guarantee its effective use. Instructional modeling -such as showing students how to interpret and apply rubric criteria during task execution-may help students internalize performance standards and engage with rubrics in a more purposeful and efficient way. Educators should consider integrating rubric training into classroom routines, particularly in courses where rubrics play a central evaluative role.

10.5. Limitations and future directions

This study involved a comprehensive multimodal data collection process, including eye-tracking, think-aloud protocols, self-report measures, electrodermal activity, and task performance. In this paper, we focused on combining eye-tracking and verbal data to gain insight into students' engagement with rubrics. While this integration strengthens the interpretive depth of the findings, triangulating process data introduces inference challenges. Eye-tracking captures implicit attentional behavior, while think-aloud protocols reflect explicit cognitive activity, two layers that do not always align. This complexity is intrinsic to multimodal process research and underscores the need for further development of analytical frameworks that allow for more integrated and theory-driven interpretations. Future studies should continue to refine methodological tools for aligning visual and verbal data streams, possibly using real-time synchronization or dynamic task-based coding systems.

Second, although think-aloud protocols offer a rich window into students' cognitive processes, their concurrent use with eye-tracking may introduce interaction effects. Specifically, the act of verbalizing while completing a task could alter natural visual behavior, potentially adding extraneous 'noise' to the eye-tracking data. This interaction, not the ecological validity of verbal protocols per se, should be considered when interpreting gaze patterns. Future research could compare think-aloud and silent conditions to isolate such effects.

Third, limitations in sample characteristics and recruitment must be acknowledged. Our participants were undergraduate students recruited through convenience sampling, which may limit generalizability to other populations, such as younger learners or experienced professionals. Future research should replicate this design with more diverse samples and in more authentic classroom settings to examine the transferability of the findings.

Fourth, the analysis of TAPs was constrained by the low frequency of rubric-related verbalizations in certain phases, particularly Phase 4. This limited our ability to compare feedback conditions systematically using verbal data. Although we reported this transparently and avoided overinterpreting the results, future studies should consider redesigning the protocol to elicit more consistent verbal engagement in later phases, perhaps by adding targeted prompts or brief reflective questions during task transitions.

Fifth, some of the statistical analyses, particularly GLM interaction models, may have been underpowered. A priori power analyses showed that our design was not sufficiently powered to detect medium-sized interactions in a $2 \times 2 \times 4$ between-subjects framework. Additionally, for the LMM analyses, power estimation remains complex due to the lack of openly available raw datasets with similar structures (Kumle et al., 2021). We thus caution against overinterpreting null results and encourage future studies to draw on larger samples and shared datasets to estimate realistic effect sizes for similar process measures.

Lastly, the task used in this study -a descriptive landscape analysis- was relatively low in cognitive complexity. While this choice allowed us to control for prior knowledge and manage protocol consistency, it may have limited the observable impact of rubric design and feedback. Future research should extend these findings by examining rubric use in more cognitively demanding tasks, such as argumentative writing or complex problem-solving, where the value of performance-level scaffolding may be more pronounced.

11. Conclusions

Rubrics are among the most widely used instructional tools across educational levels, yet how students engage with them in real time remains poorly understood. This study provides new evidence showing that even subtle design choices, such as the order in which performance levels are presented, can meaningfully shape students' visual and cognitive engagement. Presenting the highest performance level first not

only attracted more attention but was also associated with better task outcomes, particularly when paired with feedback that guided students toward productive use of the rubric.

Our findings also emphasize that rubrics are most effective when embedded within supportive feedback practices. While rubric-based feedback increased students' attention to performance criteria, process- and product-based feedback yielded stronger performance gains, underscoring the need for feedback that helps students understand how to act on evaluative information.

By combining eye-tracking and think-aloud protocols, this study contributes methodologically to the field of educational assessment, demonstrating the value of integrating multiple process-tracing methods to capture different layers of student engagement. These findings represent a first step in a broader research agenda aimed at refining both the design and implementation of rubrics to better support student learning.

CRediT authorship contribution statement

Ernesto Panadero: Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization. **Pablo Delgado:** Writing – original draft, Supervision, Formal analysis, Data curation. **David Zamorano:** Writing – original draft, Methodology, Investigation. **Leire Pinedo:** Writing – original draft, Methodology, Investigation. **Alazne Fernández-Ortube:** Writing – original draft, Methodology, Investigation. **Lucía Barrenetxea-Mínguez:** Writing – review & editing, Resources, Investigation, Data curation.

Funding

(1) Spanish National R + D call from the Ministerio de Ciencia, Innovación y Universidades (Generación del conocimiento 2020), Reference number: PID2019-108982 GB-I00. (2) Basque Government Call for Grants to support the activities of research groups of the Basque University System (2022–2025) project reference IT1624-22. (3) Basque Country Equipment 2021 call. Project: Eye tracker. Reference: EC21_2021_1_0004. (4) Ayudas a los Agentes del Sistema Andaluz del Conocimiento para la Contratación de Personal Investigador Doctor (PAIDI DOCTOR 21; Regional Government of Andalusia, Spain) to the second author.

Declaration of competing interest

The authors declare to not having any conflict of interest regarding this manuscript. This research has been approved by the ethics committee from Comité de ética de la investigación, Universidad de Deusto. Reference: ETK-5/21–22. PI: Ernesto Panadero.

Acknowledgements

We would like to express our sincere gratitude to the Editor-in-Chief, Prof. Gert Rijlaarsdam, and to the anonymous reviewers for their thoughtful and constructive feedback throughout the review process. While peer review is always an opportunity to improve a manuscript, in this case the process was particularly enriching. The reviewers' and editor's insightful challenges and generous suggestions pushed us to deepen our analyses and present our data in a richer, more detailed, and more informative way. Both the article and our thinking as researchers have been significantly enhanced by this intellectual exchange.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2025.102168.

Data availability

The data that has been used is confidential.

References

Andrade, H. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching, 53*(1), 27–32. https://doi.org/10.3200/CTCH.53.1.27-31

Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research and Evaluation, 10*(3), 1–11. Retrieved from http://pareonline.net/getvn.asp?v=10&n=3.

Andrade, H., & Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assessment & Evaluation in Higher Education, 32*(2), 159–181. https://doi.org/10.1080/02602930600801928

Arter, J., & McTighe, J. (2000). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Corwin Press.

Bresciani, M. J., Oakleaf, M., Kolkhorst, F., Nebeker, C., Barlow, J., Duncan, K., & Hickmott, J. (2009). Examining Design and Inter-Rater Reliability of a Rubric Measuring Research Quality across Multiple Disciplines. *Practical Assessment. Research, and Evaluation, 14*(1), 12. https://doi.org/10.7275/1w3h-7k62.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Virginia, USA: ASCD.

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education, 3*(22), 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

Charters, E. (2003). The Use of Think-aloud Methods in Qualitative Research An Introduction to Think-aloud Methods. *Brock Education Journal, 12*(2). https://doi.org/10.26522/brocked.v12i2.38.

Creswell, J. W., & Plano-Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd. ed.). Sage.

Dawson, P. (2017). Assessment rubrics: Towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education*, 1–14. https://doi.org/10.1080/02602938.2015.1111294

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

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics*.

Friedman, R. S., & Förster, J. (2001). The Effects of Promotion and Prevention Cues on Creativity. *Journal of Personality and Social Psychology, 81*, 1001–1013. https://doi.org/10.1037/0022-3514.81.6.1001.

George, D., & Mallery, M. (2010). *SPSS for windows step by step: A simple guide and reference* (10th ed.). Pearson.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford: OUP.

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher, 43*(5), 253–263. https://doi.org/10.3102/0013189x14542154

Isaacson, J. J., & Stacy, A. S. (2009). Rubrics for clinical evaluation: Objectifying the subjective experience. *Nurse Education in Practice, 9*(2), 134–140.

Jarodzka, H., Holmqvist, K., & Gruber, H. (2017). Eye tracking in educational science: Theoretical frameworks and research agendas. *Journal of Eye Movement Research, 10*(1), 1–18.

Jarodzka, H., Skuballa, I., & Gruber, H. (2021). Eye-Tracking in Educational Practice: Investigating Visual Perception Underlying Teaching and Learning in the Classroom. *Educ Psychol Rev, 33*, 1–10. https://doi.org/10.1007/s10648-020-09565-7.

Jones, L., Allen, B., Dunn, P., & Brooker, L. (2017). Demystifying the rubric: A five-step pedagogy to improve student understanding and utilisation of marking criteria. *Higher Education Research and Development, 36*(1), 129–142. https://doi.org/10.1080/07294360.2016.1177000

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*, 130–144.

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

Kumle, L., Võ, M. L. H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods, 53*, 2528–2543. https://doi.org/10.3758/s13428-021-01546-0

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria:what did they really say? *Organizational Research Methods, 9*(2), 202–220. https://doi.org/10.1177/1094428105284919

Landis, JR, & Koch, GG (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. PMID: 843571.

Lipnevich, A. A., Panadero, E., & Calistro, T. (2022). Unraveling the effects of rubrics and exemplars on student writing performance. *Journal of Experimental Psychology: Applied, 29*, 136. https://doi.org/10.1037/xap0000434

Loewenthal, K. M., & Lewis, C. A. (2020). *An introduction to psychological tests and scales*. Routledge.

McDonald, R. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

McDonald, S., & Petrie, H. (2013). The effect of global instructions on think-aloud testing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)* (pp. 2941–2944). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2470654.2481407.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research and Evaluation, 7*(25). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=25.

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research and Evaluation, 7*(3). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=3.

Panadero, E. (2023). Toward a paradigm shift in feedback research: Five further steps influenced by self-regulated learning theory. *Educational Psychologist, 58*(3), 193–204. https://doi.org/10.1080/00461520.2023.2223642

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. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*(0), 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., Jonsson, A., Pinedo, L., et al. (2023). Effects of Rubrics on Academic Performance, Self-Regulated Learning, and self-Efficacy: a Meta-analytic Review. *Educ Psychol Rev, 35*, 113. https://doi.org/10.1007/s10648-023-09823-4.

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.

Papageorgiou, S., Xi, X., Morgan, R., & So, Y. (2015). Developing and Validating Band Levels and Descriptors for Reporting Overall Examinee Performance. *Language Assessment Quarterly, 12*(2), 153–177. https://doi.org/10.1080/15434303.2015.1008480.

Postmes, L., Bouwmeester, R., de Kleijn, R., & van der Schaaf, M. F. (2023). Supervisors' untrained postgraduate rubric use for formative and summative purposes. *Assessment & Evaluation in Higher Education, 48*(1), 41–55. https://doi.org/10.1080/02602938.2021.2021390

Sáiz-Manzanares, M. C., Cuesta Segura, I. I., Alegre Calderon, J. M., & Peñacoba Antona, L. (2017). Effects of different types of rubric-based feedback on learning outcomes. *Frontiers in Education, 2*(34). https://doi.org/10.3389/feduc.2017.00034

Salthouse, T. A., & Ellis, C. L. (1980). Determinants of eye-fixation duration. *American Journal of Psychology, 93*, 207–234. https://doi.org/10.2307/1422228

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology, 7*, 769. https://doi.org/10.3389/fpsyg.2016.00769

Turley, E. D., & Gallagher, C. W. (2008). On the "uses" of rubrics: Reframing the great rubric debate. *English Journal, 97*(4), 87–92. https://doi.org/10.2307/30047253

van Gog, T., & Scheiter, K. (2010). Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction, 20*(2), 95–99.

Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing, 25*, 38–54. https://doi.org/10.1016/j.asw.2015.05.002

Wollenschläger, M., Hattie, J., Machts, N., Möller, J., & Harms, U. (2016). What makes rubrics effective in teacher-feedback? Transparency of learning goals is not enough. *Contemporary Educational Psychology, 44-45*, 1–11. https://doi.org/10.1016/j.cedpsych.2015.11.003