Journal of Inclusive Methodology and Technology in Learning and Teaching
ISSN 2785-5104
Anno 6, N. 3 (2026)

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

Clearer, not necessarily more accurate: generative artificial intelligence, exemplars, and revision of assessment rubrics in initial teacher education

Beatrice Doria [1,*], Valentina Grion [1], Giorgia Slaviero [2], Federica Picasso [1] and Anna Dipace [1]

[1] Department of Humanities, Education and Sport Sciences, Pegaso Telematic University, Naples, Italy; beatrice.doria@unipegaso.it; valentina.grion@unipegaso.it; federica.picasso@unipegaso.it; anna.dipace@unipegaso.it
[2] Department of Philosophy, Sociology, Education and Applied Psychology, University of Padua, Padua, Italy; giorgia.slaviero.1@phd.unipd.it
* Correspondence: beatrice.doria@unipegaso.it

Abstract: Constructing and revising an assessment rubric requires pre-service teachers to articulate a conception of quality, and revision may act either on the formulation of the descriptors or on the design of the underlying assessment framework. This study examines how pre-service primary teachers revised an analytic rubric with the support of exemplars and a generative artificial intelligence (GenAI) tool, within an authentic task. In a single-group pre–post design, two raters assessed 183 pairs of initial and final rubrics on an overall score and on nine criteria related to formulation (descriptor language) and design (rubric structure and assessment content); students indicated the resource they found most useful and rated the experience (n = 177). Overall quality improved in a widespread but moderate manner (Hodges–Lehmann estimate = 0.53 on a 1–10 scale; r_{rb} = .78; d_z = 0.57). Improvement varied across criteria (Time × Criterion interaction, η^2_p = .13): it was greatest for language clarity and moderate for structure and content, with structural precision and content validity remaining the weakest dimensions even after revision. Overall, 96.0% of students attributed a role to GenAI in the revision, whereas the perceived development of assessment competence was more modest and independent of the preferred resource. The findings suggest that revision supported by exemplars and GenAI acts more readily on the formulation of quality than on its design, and point to the need for task instructions that direct comparison towards the precision and validity of criteria.

Keywords: assessment rubrics; generative artificial intelligence; exemplars; evaluative judgement; teacher education

Copyright: © 2026 by the authors. Submitted for possible open-access publication under the terms and conditions of the Creative Commons Attribution (CC BY) licence (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The ability to recognise and construct the quality of a piece of work lies at the heart of assessment research: students need to come to hold a concept of quality comparable to that of experts and to apply it to their own work (Sadler, 1989, 2010), thereby developing the evaluative judgement that sustainable assessment regards as essential for continuing to learn beyond formal education (Boud & Soler, 2016; Grion & Serbati, 2019; Tai et al., 2018). For pre-service teachers, this ability is doubly relevant, since they are called upon to exercise it as students and, subsequently, to design the instruments by which others will be assessed. The spread

of generative artificial intelligence (GenAI) raises a new question: tools capable of rephrasing a text on demand, fluently and plausibly, may improve a piece of work without it being clear which aspect of quality they have acted upon (Bearman et al., 2024). The present study addresses this question within a circumscribed domain, namely the revision of assessment rubrics by pre-service primary teachers, observing not the development of evaluative judgement as such, but the improvement profile of the artefacts produced.

1.1. The rubric as an artefact that makes a conception of quality observable

Rubrics make explicit what constitutes quality by articulating complex competences into observable dimensions, criteria and performance levels (Brookhart, 2013; Taylor et al., 2024) and, when used formatively, they support self-assessment, self-regulation and evaluative judgement (Panadero & Jonsson, 2013; Panadero et al., 2012; Tai et al., 2018). Involving students in the definition of criteria has been associated with improvements in self-regulated learning and performance (Fraile et al., 2017). Constructing a rubric therefore places students in the position of articulating quality rather than simply receiving it, and makes their conception of quality observable in an artefact. For pre-service teachers, this is moreover an authentic task, as it reproduces an activity they will carry out in their professional practice (Villarroel et al., 2018).

The quality of a rubric, however, is not a unitary attribute. It is useful to distinguish between the *formulation* of quality, namely the clarity, accuracy and appropriateness of the language in which the descriptors are written, and its *design*, namely the structure of the rubric and its assessment content: the relevance and validity of the criteria with respect to the task, their precision, and the differentiation and progression of levels. The two components require different operations. A rubric can become more readable through rewording, whereas improving its design requires judging the coherence between criteria, task and intended learning outcomes, that is, precisely the tacit dimension of standards which, according to Sadler (2010), explicit criteria rarely convey and which develops through the direct experience of evaluating. The quality of an artefact is therefore a necessary, but not sufficient, indicator of the evaluative judgement of the person who produced it: analysing improvements in the formulation and design of rubrics does not directly measure such judgement, but makes it possible to observe which aspects of quality revision acts upon.

1.2. Exemplars and GenAI as resources for revision

From the perspective of internal feedback, students improve their own work by comparing it with external information, such as criteria, exemplars, peers' work or comments, and the formative value of any resource depends on whether such comparisons are made explicit and translated into action (Nicol, 2021; Nicol & McCallum, 2022). Exemplars, particularly when discussed dialogically, make tacit standards visible by showing concretely what quality work looks like (Carless & Chan, 2017; Tomazin et al., 2026); they therefore lend themselves to supporting comparisons that concern the underlying design of a rubric, and not only its writing.

GenAI output can also serve as a point of comparison. AI-generated feedback supports immediate and iterative revision (Yang & Niu, 2026), can be structured into feed-up and feed-forward information (Dai et al., 2024; Stephenson et al., 2026) and proves most useful when prompts refer to explicit criteria, such as rubrics

(Heinrich et al., 2025; Mills et al., 2025; Pan, 2025), ideally within human-in-the-loop models in which the teacher retains a mediating role (Cooper et al., 2026; Lim & Go, 2025; Picasso et al., 2025). The same features, however, entail a specific risk: since GenAI produces fluent and plausible text on demand, revision may concentrate on rewording, which GenAI performs with particular ease, while leaving the design unchanged. For this reason, Bearman et al. (2024) argue that GenAI makes evaluative judgement more, not less, necessary: students must learn to judge the quality of GenAI outputs and processes.

This yields a testable expectation regarding the artefacts. If revision acts mainly on formulation, improvements should be concentrated in the language of the descriptors; if it also involves a rethinking of the underlying assessment design, they should extend to the structure and content of the rubric. In an activity that combines GenAI and exemplars, change cannot be attributed to a single resource, but the improvement profile across rubric dimensions indicates which component of quality the revision, as a whole, has acted upon.

1.3. Perceptions and performance

A previous study of the same learning experience (Doria et al., 2026) found that, after a GenAI-supported rubric revision activity, pre-service teachers reported more reflective and criteria-oriented approaches, more positive attitudes towards AI and lower technostress, with no differences according to the resource perceived as most useful. These findings rest entirely on self-reported data, and the authors themselves called for the use of performance-based measures. Students' perceptions of the usefulness of a resource do not necessarily coincide with the effect of that resource on the work produced: complementing the survey of perceptions with an independent assessment of the artefacts makes it possible to verify whether, and on which dimensions, perceived improvement is reflected in the products.

1.4. The present study

The present study examines an activity in which pre-service primary teachers constructed an analytic rubric and revised it with the support of three exemplars provided by the instructor and of a GenAI tool. The initial and final rubrics were assessed by two raters, while students indicated the resource they found most useful and rated the learning experience.

The study concerns the same cohort analysed by Doria et al. (2026), but draws on different data: Doria et al. (2026) analysed the responses of the 144 students who had completed both the initial and the final questionnaire, whereas the present study analyses the rubrics produced by all participants (183 pairs) and a set of post-intervention items, to which 177 students responded. Its contribution lies in offering an external assessment of the artefacts, which makes it possible to describe the dimensions on which revision produced improvements, and in comparing, at group level, the resulting picture with students' perceptions.

The study is guided by three research questions:

RQ1. To what extent does the overall quality of the rubrics improve after revision supported by exemplars and GenAI?

RQ2. Is the improvement concentrated on the formulation of quality (descriptor language), or does it extend to its design (rubric structure and assessment content)?

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

RQ3. Which resource (exemplars, GenAI or both) is perceived as most useful for revision, and does the evaluation of the learning experience, including the perceived development of assessment competences, vary according to the preferred resource?

2. Materials and Methods

2.1. Research design and participants

The study adopted a single-group pre–post design, embedded in a workshop activity within the five-year single-cycle master's degree programme in Primary Education at a university in northern Italy. A total of 183 students took part, each of whom produced an initial and a final version of their own rubric; 177 of them also responded to the post-intervention items.

2.2. Instructional intervention

The activity, carried out individually in self-paced mode on Moodle and without summative purposes, was conducted in several phases.

In the first phase, students drafted a rubric by choosing a hypothetical competence to be assessed, an instructional context, the assessment criteria for the competence and four proficiency levels.

After submitting the first version of the rubric, students revised it using three exemplars (on the combined use of rubrics and annotated exemplars, see Tomazin et al., 2026) and a GenAI tool of their choice in order to identify areas for improvement in their products.

Subsequently, each student was required to assess the final version of their own rubric. The activity concluded with the final questionnaire, aimed at evaluating students' perceptions of the tools used and of the learning experience as a whole.

2.3. Instruments and data sources

2.3.1. Questionnaire

The first data source consists of the items administered at the end of the activity. A closed-ended item asked which resource students had found most useful for improving the rubric (GenAI, exemplars, both, neither). In addition, five items on a six-point Likert scale measured the degree of agreement with the following statements: 1. “Using AI to improve my rubric was a valuable learning experience”; 2. “The exemplars helped me more than AI to understand how to improve the rubric”; 3. “During the activity I developed my assessment competences”; 4. “I think that in the future I could use AI tools to support assessment and instructional design”; 5. “Reflecting on the different tools (exemplars and AI) helped me to better understand the assessment process”.

2.3.2. Analytic rubrics

The second data source consists of the rubrics produced before and after revision; the two versions of each rubric constitute paired observations.

2.4. Assessment of the artefacts

The initial and final rubrics were assessed by two independent raters using an analytic grid designed to capture their pedagogical, structural and linguistic quality.

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

The grid comprises nine criteria grouped into three macro-areas, namely rubric structure, descriptor language and assessment content (Table 1), each rated on an ordinal scale from 1 to 4; an overall rubric score (from 1 to 10) was also assigned. The raters were not blind to the version, initial or final, of the rubrics assessed.

Table 1. Structure of the rubric assessment grid.

Macro-area	Criteria
Rubric structure	Clarity; Precision; Coherence
Descriptor language	Clarity; Accuracy; Appropriateness
Assessment content	Precision; Coherence; Validity

Note. Each criterion is rated on a scale from 1 to 4; the overall rubric score is expressed on a scale from 1 to 10.

To distinguish substantive improvements from purely formal changes, the raters considered, in addition to linguistic correctness, the relevance of the criteria, alignment with the task, the clarity and observability of the descriptors, the differentiation and progression between levels, and the completeness and usability of the rubric for formative purposes. Consistent with the distinction introduced in Section 1.1, the descriptor language macro-area was considered indicative of the formulation of quality, which is more amenable to rewording, whereas the rubric structure and assessment content macro-areas were considered indicative of its design, which requires a revision of the underlying assessment framework. This distinction provides the interpretive key for answering RQ2.

2.5. Data analysis

Analyses were conducted with jamovi 2.6 (The jamovi project, 2024), based on R 4.4 (R Core Team, 2024); the repeated-measures ANOVA was estimated with the afex package (Singmann, 2023), and marginal means and post hoc comparisons with emmeans (Lenth, 2023). Inter-rater agreement on the overall score was estimated using intraclass correlation coefficients for absolute agreement, ICC(A,1) and ICC(A,*k*) (McGraw & Wong, 1996), interpreted according to Koo and Li (2016); the inter-rater correlation and the mean difference were also calculated, and subsequent analyses were conducted on the mean of the two scores.

To answer RQ1 and RQ2, the rubrics were analysed; to answer RQ3, the post-intervention items were analysed.

For RQ1, since the Shapiro–Wilk test indicated a non-normal distribution of the paired differences, the Wilcoxon signed-rank test was adopted as the primary analysis, with the Hodges–Lehmann estimate, its 95% confidence interval and the rank-biserial correlation as the effect size measure (Kerby, 2014). The paired-samples *t* test, with mean difference, confidence interval and Cohen's *d* for repeated measures (d_z; Lakens, 2013), was used as a convergent check. All differences are oriented in the post − pre direction.

For RQ2, the same tests were applied to each criterion, with Holm (1979) correction, recording the number of rubrics with a non-zero difference in order to assess how widespread the change was. A 2 (Time) × 9 (Criterion) repeated-measures ANOVA tested whether the magnitude of improvement differed across criteria, with Greenhouse–Geisser corrections where sphericity was violated and Tukey-corrected post hoc comparisons on the estimated marginal means.

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

For RQ3, frequencies for the most useful resource and descriptive statistics for the Likert items were calculated for the whole sample and by preferred resource; the two numerically substantial groups (GenAI and Both) were compared using the Mann–Whitney *U* test, the rank-biserial correlation and Holm correction.

3. Results

3.1. Inter-rater agreement

Agreement between the two raters on the overall score was good, with ICC(A,1) = .750 and ICC(A,*k*) = .857 (Koo & Li, 2016), although the single-rater index lies at the lower bound of this range; the higher ICC(A,*k*) value supports the use of the mean of the two ratings. The inter-rater correlation was high (r = .799), indicating good consistency in the ranking of the rubrics. Rater 1 assigned slightly higher scores than Rater 2 (mean difference ≈ 0.31 points): this systematic deviation, modest relative to the range of the scale, accounts for the difference between correlation and absolute agreement and does not compromise the pre–post comparison, since it affects both time points equally.

3.2. Overall rubric quality before and after revision (RQ1)

The mean overall score increased from the initial version (M = 7.66, Mdn = 7.85, SD = 1.36) to the final version (M = 8.20, Mdn = 8.55, SD = 1.21; Table 2). The rise in central tendency was accompanied by a slight reduction in variability and an increase in the minimum value (from 2.40 to 3.05), indicating a moderate upward homogenisation of quality, although at both time points a small group of rubrics with markedly lower scores remained (Figure 1).

Table 2. Descriptive statistics for the overall score before and after the intervention.

	N	*M*	*Mdn*	*SD*	Min	Max
Pre	182	7.66	7.85	1.36	2.40	9.85
Post	182	8.20	8.55	1.21	3.05	10.00

Note. The overall score is unavailable for one of the 183 rubrics.

The distribution of the paired differences deviates significantly from normality, W = .809, p < .001. The Wilcoxon signed-rank test indicates a significant improvement from the initial to the final version, W = 1364.0, p < .001, with a Hodges–Lehmann estimate of 0.53 points (95% CI [0.43, 0.65]) and a large effect (r_{rb} = .78). The paired-samples t test leads to a convergent conclusion, $t(181)$ = 7.71, p < .001, mean difference = 0.53, 95% CI [0.40, 0.67], d_z = 0.57, corresponding to a medium effect (Table 3). Of the 182 rubrics, 156 (85.7%) showed a change in the overall score, while 26 remained unchanged.

Table 3. Pre–post comparison of the overall score.

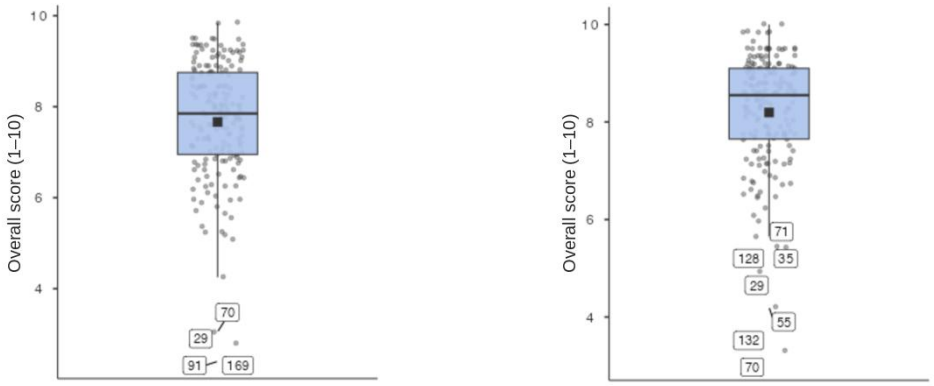
Test	Statistic	*df*	*p*	Estimated difference	95% CI	Effect size
Paired-samples *t* test	t = 7.71	181	< .001	0.531	[0.395, 0.667]	d_z = 0.571
Wilcoxon	W = 1364.0	–	< .001	0.525	[0.425, 0.650]	r_{rb} = 0.777

Note. N = 182. Differences are oriented in the post − pre direction: positive values indicate improvement. For the t test, the mean difference and Cohen's d for repeated measures (d_z) are reported; for the Wilcoxon test, the Hodges–Lehmann estimate and the rank-

biserial correlation (r_{rb}). *W* corresponds to the sum of the ranks of negative differences (deteriorations), calculated excluding the 26 pairs with zero difference. CI = confidence interval.

The two effect size indices differ in nature: the rank-biserial correlation, calculated only on pairs with a non-zero difference, expresses the consistency of the direction of change and signals a clear predominance of improvements, whereas d_z relates the mean magnitude of change to its variability, which is inflated by the extreme cases visible in Figure 1. The improvement, although modest in absolute terms (about half a point on a ten-point scale), was therefore widespread across rubrics.

Figure 1. Distribution of the overall score in the initial (**a**) and final (**b**) versions of the rubrics. The black square indicates the mean; numerical labels identify outliers.



(**a**) (**b**)

3.3. Variation in change across rubric quality criteria (RQ2)

3.3.1. Pre–post comparisons by criterion

Table 4 and Figure 2 report the pre–post comparisons for each of the nine criteria. Here too, the differences deviate from normality (*W* between .477 and .826, all $p < .001$). The two tests lead to concordant conclusions: eight criteria show a significant improvement and one a significant decrease, and all comparisons remain significant after Holm correction (adjusted $p \leq .028$ for the *t* test and $\leq .022$ for the Wilcoxon test). The magnitude of change, however, varies considerably across dimensions, outlining three distinct profiles.

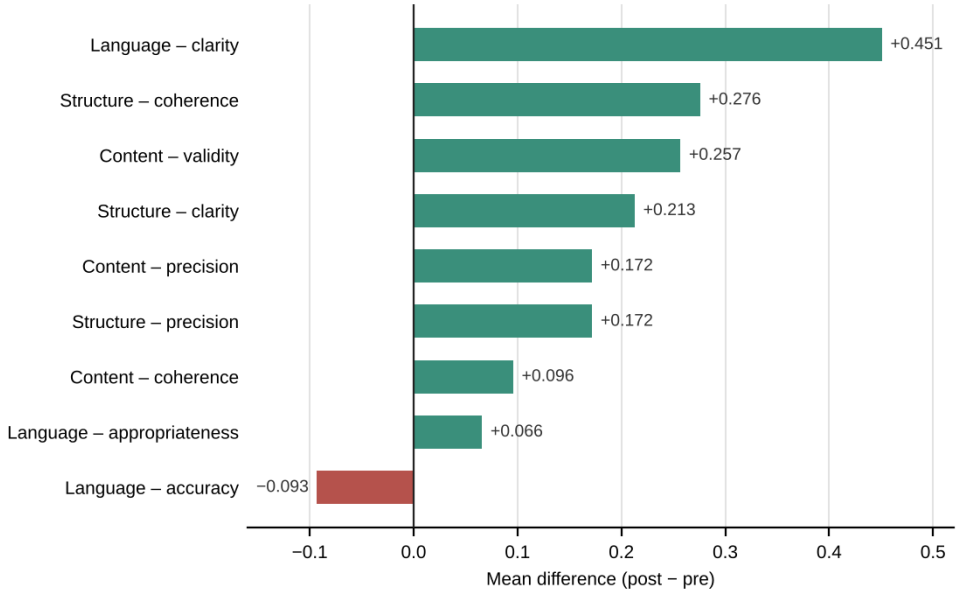
EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

Table 4. Pre–post comparisons by criterion.

			t test				Wilcoxon test			
Criterion	Pre *M* (*SD*)	Post *M* (*SD*)	Δ [95% CI]	*t*(182)	*p*	d_z	*W*	*p*	r_{rb}	$n \neq 0$
Structure – clarity	3.68 (0.50)	3.89 (0.26)	+0.21 [0.15, 0.28]	6.29	< .001	0.47	39.0	< .001	0.95	54
Structure – precision	2.57 (0.49)	2.74 (0.48)	+0.17 [0.11, 0.23]	5.90	< .001	0.44	222.5	< .001	0.77	61
Structure – coherence	3.01 (0.86)	3.29 (0.74)	+0.28 [0.19, 0.36]	6.62	< .001	0.49	243.0	< .001	0.83	75
Language – clarity	3.13 (0.79)	3.58 (0.58)	+0.45 [0.35, 0.55]	9.25	< .001	0.68	219.5	< .001	0.91	99
Language – accuracy	3.46 (0.97)	3.37 (1.10)	−0.09 [−0.17, −0.02]	−2.41	.017	−0.18	228.0	.022	−0.52	24
Language – appropriateness	3.80 (0.46)	3.86 (0.44)	+0.07 [0.01, 0.12]	2.48	.014	0.18	137.0	.007	0.51	33
Content – precision	3.16 (0.50)	3.33 (0.46)	+0.17 [0.12, 0.23]	6.06	< .001	0.45	196.0	< .001	0.80	62
Content – coherence	3.71 (0.52)	3.81 (0.48)	+0.10 [0.03, 0.16]	2.85	.005	0.21	219.0	.005	0.49	41
Content – validity	2.74 (0.83)	3.00 (0.78)	+0.26 [0.17, 0.34]	5.86	< .001	0.43	647.5	< .001	0.68	90

Note. $N = 183$. Δ = mean post − pre difference with 95% confidence interval. The sign of *t*, d_z and r_{rb} is oriented in the post − pre direction (positive values indicate improvement); *W* corresponds to the sum of the ranks of negative differences (deteriorations), calculated only on pairs with a non-zero difference. $n \neq 0$ = number of rubrics with a non-zero difference. All comparisons remain significant after Holm correction (adjusted $p \leq .028$ for the *t* test; ≤ .022 for the Wilcoxon test).

Figure 2. Mean post − pre difference for each criterion.



The most marked increase concerns language clarity (Δ = +0.45, 95% CI [0.35, 0.55]; $t(182) = 9.25$, $p < .001$, $d_z = 0.68$; $r_{rb} = .91$), the only criterion to reach a medium-to-large effect. A second group, with increases between +0.17 and +0.28 and small-to-medium effects ($d_z = 0.43$–0.49), comprises structural coherence, content validity, structural clarity, content precision and structural precision. Small but significant increases are observed for content coherence (Δ = +0.10, $d_z = 0.21$) and language appropriateness (Δ = +0.07, $d_z = 0.18$), both of which had initial scores already close to the scale maximum ($M = 3.71$ and 3.80; $Mdn = 4.00$) and hence limited room for improvement.

The proportion of rubrics that actually changed their score on each criterion (Table 4, last column) ranges from 13.1% for language accuracy (24 of 183 rubrics)

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

to 54.1% for language clarity (99 rubrics), and lies between 22.4% (content coherence) and 49.2% (content validity) for the structure and content criteria. The high rank-biserial correlation values on some criteria, such as structural clarity (r_{rb} = .95, with 54 rubrics changed), therefore indicate that almost all changes were improvements, not that most rubrics improved on that criterion. Language clarity thus emerges as the dimension in which change was at once the largest and the most widespread.

Language accuracy is the only criterion showing a decrease, small but statistically significant (Δ = −0.09, 95% CI [−0.17, −0.02]; $t(182) = -2.41$, $p = .017$, $d_z = -0.18$; $W = 228.0$, $p = .022$, $r_{rb} = -.52$), accompanied by an increase in variability in the final version (SD from 0.97 to 1.10). This finding, however, concerns a limited number of rubrics: only 24 changed their score on this criterion, predominantly for the worse. Moreover, the two criteria with the lowest initial scores, structural precision (M = 2.57) and content validity (M = 2.74), although improving significantly, remain the dimensions with the lowest scores even after revision (M = 2.74 and 3.00).

3.3.2. Time × Criterion interaction

The 2 × 9 repeated-measures ANOVA (Table 5) revealed a significant main effect of Time, $F(1, 182) = 53.4$, $p < .001$, $\eta^2_p = .23$, confirming an overall increase in mean scores across criteria from pre (M = 3.25, 95% CI [3.19, 3.32]) to post (M = 3.43, 95% CI [3.37, 3.49]), corresponding to a difference of 0.18 points (SE = 0.025, $t(182) = 7.31$, $p_{Tukey} < .001$). Since Mauchly's test indicated a violation of sphericity for the Criterion factor and for the interaction (see the note to Table 5), Greenhouse–Geisser corrected degrees of freedom are reported for these effects.

Table 5. 2 (Time) × 9 (Criterion) repeated-measures ANOVA.

Effect	*SS*	*df*	*MS*	*F*	*p*	η^2_p
Time	26.3	1	26.33	53.4	< .001	.227
Residual	89.8	182	0.49			
Criterion	497.5	8	62.19	122.9	< .001	.403
Residual	736.5	1456	0.51			
Time × Criterion	16.9	8	2.11	27.3	< .001	.130
Residual	112.4	1456	0.08			

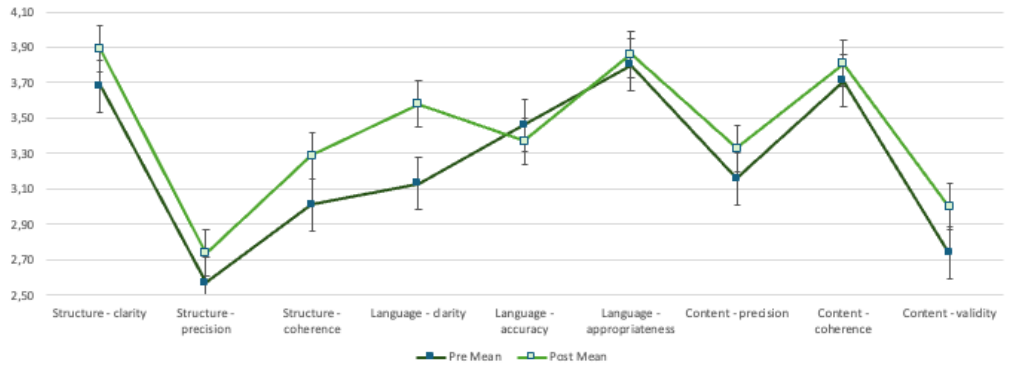
Note. Type III sums of squares; uncorrected degrees of freedom are reported. Mauchly's test of sphericity: Criterion, $W = 0.003$, $p < .001$, $\varepsilon_{GG} = 0.441$, $\varepsilon_{HF} = 0.451$; Time × Criterion, $W = 0.056$, $p < .001$, $\varepsilon_{GG} = 0.582$, $\varepsilon_{HF} = 0.599$; for Time (two levels), sphericity is satisfied by definition (ε = 1.000). The Greenhouse–Geisser corrected degrees of freedom are $F(3.53, 642.10)$ for Criterion and $F(4.66, 847.39)$ for Time × Criterion. GG = Greenhouse–Geisser; HF = Huynh–Feldt.

A significant main effect of Criterion emerged, $F(3.53, 642.10) = 122.9$, $p < .001$, $\eta^2_p = .40$, indicating systematic differences in level across the rubric dimensions. Considering the marginal means across the two time points, the highest scores were observed for language appropriateness (M = 3.83), structural clarity (M = 3.79) and content coherence (M = 3.76), and the lowest for structural precision (M = 2.65) and content validity (M = 2.87); Tukey-corrected post hoc comparisons indicated significant differences in 29 of the 36 pairwise comparisons between criteria.

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

Finally, the Time × Criterion interaction was significant, $F(4.66, 847.39) = 27.3$, $p < .001$, $\eta^2_p = .13$, formally confirming that the improvement was not uniform but varied by criterion. The estimated marginal means (Figure 3 and Table 4) allow this interaction to be qualified. The profile of the criteria remains substantially stable across the two time points, with the same dimensions at the highest and lowest levels; the divergence between the two profiles is greatest for language clarity, whereas language accuracy is the only point at which the post profile falls below the pre profile. Pre–post post hoc comparisons within each criterion, with the more conservative Tukey correction applied to the full set of 153 pairwise comparisons, confirm the significance of change for six criteria (structural clarity, precision and coherence; language clarity; content precision and validity; all $p_{Tukey} < .001$), whereas the changes in language appropriateness ($p_{Tukey} = .546$), content coherence ($p_{Tukey} = .294$) and language accuracy ($p_{Tukey} = .600$) are not significant. These changes, small in magnitude and confined to a minority of rubrics, should therefore be interpreted with caution, as they are sensitive to the correction procedure adopted.

Figure 3. Estimated marginal means pre and post by criterion. Bars indicate the standard error.



3.4. Resource perceived as most useful and evaluation of the experience (RQ3)

Of the 177 respondents, 96 (54.2%) indicated GenAI as the most useful resource for improving the rubric, 74 (41.8%) the combination of GenAI and exemplars, 6 (3.4%) exemplars alone and 1 (0.6%) neither (Table 6). Overall, therefore, 96.0% of participants attributed a role to GenAI in the revision process, whereas exemplars were recognised as useful by 45.2%, almost exclusively in combination with AI. The two resources thus do not appear to be perceived as alternatives: the choice of exemplars alone is marginal, while a substantial proportion of participants recognise their value alongside that of AI.

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

Table 6. Resource perceived as most useful for improving the rubric.

Resource	*n*	%
Generative artificial intelligence	96	54.2
Both (GenAI and exemplars)	74	41.8
Exemplars	6	3.4
Neither	1	0.6
Total	177	100.0

Note. N = 177 respondents to the item "Which tool did you find most useful for improving your rubric?".

Across the whole sample (Table 7), the item concerning AI as a learning experience obtained the highest mean (M = 5.15, SD = 0.86), with 80.2% of responses in the two highest scale points, followed by the intention to use AI in the future (M = 4.90, SD = 0.92; 73.4% at points 5–6) and by the usefulness of reflecting on the tools (M = 4.84, SD = 0.91; 68.9%). The perceived development of assessment competences showed a lower mean (M = 4.58, SD = 0.86; 58.2% at points 5–6), whereas the item stating that the exemplars had been more useful than AI fell below the scale midpoint (M = 3.15, SD = 1.41), with the highest variability and 37.3% of responses in the agreement range (4–6).

Table 7. Mean scores on the post-intervention items by resource perceived as most useful.

Item	Total (N = 177)	GenAI (n = 96)	Both (n = 74)	Exemplars (n = 6)	U	p	r_{rb}
1. AI as a learning experience	5.15 (0.86)	5.12 (0.89)	5.27 (0.75)	4.17 (1.17)	3298.5	.392	.07
2. Exemplars more useful than AI	3.15 (1.41)	2.61 (1.39)	3.62 (1.03)	5.83 (0.41)	1919.0	< .001	.46
3. Development of assessment competences	4.58 (0.86)	4.58 (0.88)	4.58 (0.79)	4.83 (1.17)	3510.5	.889	.01
4. Future use of AI	4.90 (0.92)	4.93 (0.95)	5.01 (0.75)	3.33 (0.82)	3496.0	.850	.02
5. Usefulness of reflecting on the tools	4.84 (0.91)	4.80 (0.95)	4.91 (0.80)	4.83 (1.47)	3402.5	.617	.04

Note. Values are expressed as M (SD) on a scale from 1 to 6. U, p and r_{rb} refer to the comparison between the GenAI and Both groups (Mann–Whitney test); positive r_{rb} values indicate higher scores in the Both group. After Holm correction, only the comparison concerning exemplars remains significant (adjusted p < .001; adjusted p = 1.000 for the other items). The Exemplars group is reported for descriptive purposes; the respondent who indicated no resource (n = 1) is excluded from the groups but included in the total.

The comparison between the GenAI and Both groups reveals only one significant difference, concerning the greater usefulness of exemplars compared with AI (U = 1919.0, p < .001, Holm-adjusted p < .001, r_{rb} = .46), with higher values in the Both group (M = 3.62, SD = 1.03) than in the GenAI group (M = 2.61, SD = 1.39). For the remaining four items, differences are not significant (all $p \geq .392$) and effect sizes are negligible ($r_{rb} \leq .07$). The Both group does not express a less favourable evaluation of AI, either as a learning experience (M = 5.27 vs 5.12) or in terms of intention of future use (M = 5.01 vs 4.93), outlining an integrative stance in which the value attributed to AI is not diminished but complemented by the recognition of concrete examples. The small group that indicated exemplars alone shows the opposite profile, with strong agreement on the greater usefulness of exemplars (M = 5.83, SD = 0.41) and lower values for AI as a learning experience (M = 4.17, SD = 1.17) and for intention of future use (M = 3.33, SD = 0.82); given

the group size (n = 6), these values should be regarded as purely indicative. By contrast, the perceived development of assessment competences (M = 4.58, 4.58 and 4.83) and the usefulness attributed to reflecting on the tools (M = 4.80, 4.91 and 4.83) show substantially overlapping values across the three groups: the perception of having developed assessment competences appears independent of the preferred resource and attributable to the experience as a whole rather than to any specific form of support.

4. Discussion

Rubric quality improved after revision, both in the overall score and on most criteria. The improvement was widespread, since it affected 85.7% of the rubrics, but moderate, amounting to about half a point on a ten-point scale. Consistent with the single-group design, it should be read as an effect of the activity as a whole, which combined explicit criteria, exemplars and GenAI, and not as an effect of GenAI alone; it should also be read as an improvement in the quality of a product, not as direct evidence of growth in evaluative judgement. Compared with Doria et al. (2026), the present study complements students' perceptions with an external assessment of the artefacts, which confirms an improvement and specifies its distribution across the rubric dimensions.

The improvement profile is the most informative finding. The largest and most widespread change concerns language clarity, a dimension of formulation. Structure and assessment content also improve, with small-to-medium effects, so that revision was not confined to the linguistic surface; however, structural precision and content validity, the dimensions most directly linked to design, started from the lowest levels and remain the weakest even after revision. In Sadler's (2010) terms, appropriating a concept of quality involves more than recognising well-formulated descriptions: it requires judging the coherence between criteria, task and intended learning outcomes, which is precisely what precision and validity capture. The observed profile is therefore consistent with the expectation formulated in the introduction: revision acted more readily on the formulation of quality than on its design.

This profile admits two readings. In the first, students used the exemplars and the GenAI output as points of comparison, generating internal feedback that led them to refine the formulation of the criteria (Nicol, 2021); in the second, they adopted reworded descriptors with limited scrutiny, so that the rubric became clearer without any change in the underlying conception of quality, amounting to the very delegation of judgement against which Bearman et al. (2024) warn. The design of the study allows neither the two readings to be distinguished nor the improvements to be attributed to GenAI rather than to the exemplars. The slight decline in language accuracy, confined to 24 rubrics and not significant under Tukey correction, does not permit any conclusions; it is compatible with the hypothesis, yet to be tested, that rewording may at times move descriptors away from appropriate pedagogical and disciplinary terminology.

Students' perceptions, considered at group level, are consistent with this picture. Almost all of them attributed a role to GenAI in improving the rubric (96.0%), whereas only 3.4% indicated exemplars alone, and the experience with AI was rated as highly formative. The perceived development of assessment competences, by contrast, received the lowest mean rating among the items expressing agreement with the experience and did not vary according to the

preferred resource, in line with Doria et al. (2026). In other words, students recognise GenAI as the main source of improvement, but attribute to the experience only moderate growth in their own assessment competence, just as the artefacts show a more marked improvement in formulation than in design. The group that valued both GenAI and exemplars did not rate AI less favourably, but explicitly acknowledged the contribution of concrete models of quality. Since perceptions and artefacts were not linked at the individual level, this correspondence should be understood as a convergence between two aggregate pictures and not as an association between the perceptions and the performance of individual students.

From the perspective of sustainable assessment (Boud & Soler, 2016), the implication for initial teacher education is that resources for revision should be designed so as to direct comparison towards the design of quality, and not only towards its formulation. Since the formative potential of AI feedback depends on prompt design and on the presence of explicit criteria (Heinrich et al., 2025; Mills et al., 2025; Pan, 2025), task instructions can ask students to query GenAI about the precision, coherence and validity of the criteria, as well as about the clarity of the descriptors. They can also ask students to indicate which suggestions they accepted, modified or rejected and why, to compare their own rubric dialogically with the exemplars (Carless & Chan, 2017) and with the intended learning outcomes, and to negotiate criteria before turning to GenAI (Fraile et al., 2017), within human-in-the-loop models in which teachers retain responsibility for mediation (Cooper et al., 2026; Lim & Go, 2025). Designs of this kind turn AI output into an object of judgement rather than a substitute for it, an aspect of twofold relevance for pre-service teachers, who will need to cultivate the same capacity in their own pupils.

The study has several limitations. The single-group pre–post design, in which GenAI and exemplars were used jointly, does not allow improvements to be attributed to a specific resource, and the outcome variable, rubric quality, is an indirect indicator of evaluative judgement. The raters were not blind to the version of the rubrics, initial or final, which may have favoured the assignment of higher scores to the revised versions; moreover, inter-rater agreement was estimated on the overall score only. Perceptions were measured through single items rather than validated scales and were not linked to changes in the artefacts at the individual level. The intervention took place in a single context, over a short period and with non-prescribed use of GenAI, which increased authenticity but reduced control over prompting strategies; finally, the cohort is the same as that analysed by Doria et al. (2026). Future research should combine comparative designs, blind assessment of artefacts and process data, such as prompts and justifications for changes, in order to establish whether the observed improvements reflect developing evaluative judgement.

5. Conclusions

Embedded in an authentic task of rubric construction and revision, revision supported by exemplars and GenAI improved the quality of pre-service teachers' rubrics in a widespread but moderate manner, acting more readily on the formulation of quality than on its design: the rubrics became, above all, clearer, whereas precision and validity remained the weakest dimensions. Students attributed a central role in the improvement to GenAI, but perceived more limited growth in their own assessment competence. The message of the study, therefore, is not that

GenAI improves or worsens assessment design, but rather that, in the absence of targeted task instructions, revision risks concentrating on what is easiest to reword. Integrating AI into teacher education thus requires tasks that shift comparison from the clarity of the descriptors to the precision and validity of the criteria, together with explicit moments in which students account for the choices they have made.

References

Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education*. Advance online publication. https://doi.org/10.1080/02602938.2024.2335321

Boud, D., & Soler, R. (2016). Sustainable assessment revisited. *Assessment & Evaluation in Higher Education*, *41*(3), 400–413. https://doi.org/10.1080/02602938.2015.1018133

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

Carless, D., & Chan, K. K. H. (2017). Managing dialogic use of exemplars. *Assessment & Evaluation in Higher Education*, *42*(6), 930–941. https://doi.org/10.1080/02602938.2016.1211246

Cooper, G., Ilich, K., & Sturrock, K. (2026). Using generative artificial intelligence to reimagine feedback in higher education: A collaborative autoethnography. *Assessment & Evaluation in Higher Education*, 1–15.

Dai, W., Tsai, Y.-S., Lin, J., Aldino, A., Jin, H., Li, T., . . . Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, *7*, Article 100299.

Doria, B., Slaviero, G., Zaggia, C., & Grion, V. (2026). Generative artificial intelligence as evaluative scaffolding: Effects on students' perceptions and dispositions in an authentic learning task. *The Turkish Online Journal of Educational Technology*, *25*(3), 11–29.

Fraile, J., Panadero, E., & Pardo, R. (2017). Co-creating rubrics: The effects on self-regulated learning, self-efficacy and performance of establishing assessment criteria with students. *Studies in Educational Evaluation*, *53*, 69–76. https://doi.org/10.1016/j.stueduc.2017.03.003

Grion, V., & Serbati, A. (2019). *Valutazione sostenibile e feedback nei contesti universitari* [Sustainable assessment and feedback in university contexts]. Pensa MultiMedia.

Heinrich, T., Baily, S., Chen, K. W., DeOliveira, J., Park, S., & Wang, N. C. H. (2025). AI-assisted grading and personalized feedback in large political science classes: Results from randomized controlled trials. *PLOS ONE*, *20*(8), Article e0328041.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

The jamovi project. (2024). *jamovi* (Version 2.6) [Computer software]. https://www.jamovi.org

Kerby, D. S. (2014). The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology*, *3*. https://doi.org/10.2466/11.IT.3.1

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, Article 863. https://doi.org/10.3389/fpsyg.2013.00863

Lenth, R. (2023). *emmeans: Estimated marginal means, aka least-squares means* [R package]. https://cran.r-project.org/package=emmeans

Lim, J. M., & Go, C. (2025). Integrating ChatGPT into teacher feedback: Practical insights for L2 writing instruction. *TESL-EJ*, *29*(3).

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Mills, E., Mizouri, A., & Peach, A. (2025). Prompting better feedback: A study of custom GPT for formative assessment in undergraduate physics. *Education Sciences*, *15*(8), Article 1058.

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education*, *46*(5), 756–778.

Nicol, D., & McCallum, S. (2022). Making internal feedback explicit: Exploiting the multiple comparisons that occur during peer review. *Assessment & Evaluation in Higher Education*, *47*(3), 424–443. https://doi.org/10.1080/02602938.2021.1924620

Pan, Y. (2025). Leveraging generative AI powered rubric-indexed feedback as a formative assessment strategy for enhancing medical English education. *Discover Computing*, *28*(1), Article 284.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, *9*, 129–144.

Panadero, E., Tapia, J. A., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and Individual Differences*, *22*(6), 806–813.

Picasso, F., Agostini, D., & Serbati, A. (2025). Exploring the potential of AI for authentic assessment in education: Towards a new model of interaction. In Vilalta-Perdomo, Scroccaro, Salinas-Navarro, & Michel-Villarreal (Eds.), *The Emerald handbook of active learning for authentic assessment*. Emerald. https://doi.org/10.1108/9781837978571

R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4) [Computer software]. https://cran.r-project.org

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144.

Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, *35*(5), 535–550. https://doi.org/10.1080/02602930903541015

Singmann, H. (2023). *afex: Analysis of factorial experiments* [R package]. https://cran.r-project.org/package=afex

Stephenson, E., Morrigan, K., Irfan, A., Bascombe, K., & Okorie, M. (2026). Artificial intelligence-assisted feedback in pharmacology education: A pilot evaluation of a custom generative model. *BMC Medical Education*, *26*(1), Article 1238.

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, *76*(3), 467–481. https://doi.org/10.1007/s10734-017-0220-3

Taylor, B., Kisby, F., & Reedy, A. (2024). Rubrics in higher education: An exploration of undergraduate students' understanding and perspectives. *Assessment & Evaluation in Higher Education*, *49*(6), 799–809. https://doi.org/10.1080/02602938.2023.2299330

Tomazin, L., Bentenuto, A., Perzolli, S., Venuti, P., Serbati, A., Picasso, F., & Lipnevich, A. A. (2026). Teacher feedback, rubrics, and annotated exemplars: Examining the effects on Italian psychology university students' performance on writing assignments. *Assessment & Evaluation in Higher Education*, 1–14.

Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment & Evaluation in Higher Education*, *43*(5), 840–854.

Yang, J., & Niu, T. (2026). Exploring of the impact of AI feedback on college students' critical thinking: A intervention experiment based on design discipline. *Frontiers in Psychology*, *17*, Article 1806913.