Rubrics as Scales of Graded Criteria in Educational Assessment

Despoina Styla[1], Vasileios Mellos[2], Dimitrios Mantzaris[3]

[1]Laboratory Teaching Staff, Department of Primary Education, University of Thessaly, Volos, Greece

[2]PhD Candidate Student, Department of Economics, University of Thessaly, Volos, Greece

[3]Laboratory Teaching Staff, Department of Nursing, University of Thessaly, Larissa, Greece

DOI: https://doi.org/10.51244/IJRSI.2026.1309000106

Received: 19 September 2026; Accepted: 24 September 2026; Published: 09 October 2026

ABSTRACT

Although rubrics are widely recommended in educational assessment, much of the literature on their use is based on practitioner guidance rather than empirical research, and few existing reviews have considered several outcomes or educational levels together. This systematic scoping review maps (a) the nature of the evidence on rubric use; (b) the functions that rubrics serve for students, teachers, and families. The review was conducted in accordance with the PRISMA extension for scoping reviews. Scopus, Web of Science, ERIC, and Google Scholar were searched between June and August 2026 for English- and Greek-language publications dated from January 1997 to July 2026, and the searches were supplemented by backward and forward citation searching. Of the 220 records screened and 85 full texts assessed, 60 sources met the eligibility criteria. Meta-analytic evidence indicates a moderate positive effect of rubrics on academic performance and smaller effects on self-regulated learning and self-efficacy, provided that rubrics are used formatively and students are explicitly trained to use them. Evidence on scoring reliability is mixed and depends on rater training, the availability of exemplars, and analytic design. The claim that rubrics restrict creativity is supported as well in sources, and rubrics improve the accuracy of students' self-assessment more consistently than they improve the quality of the work itself. Empirical studies are concentrated in higher education, especially teacher education, whereas online and vocational settings have received little attention. On this basis, we propose a four-function typology of rubrics (communicative, instructional, evaluative, and professional), moderated by design features, implementation conditions, and context. From this typology we derive concrete recommendations for rater training, student training, and the supported by teachers and institutions use of rubrics, together with a research agenda that focuses on conditions and mechanisms.

Keywords: Rubrics; Analytic And Holistic Rubrics; Formative Assessment; Self-Regulated Learning; Rater Training; Scoping Review.

INTRODUCTION

Rubrics, also known as scales of graded criteria, are tools that teachers use "to judge the degree of students' understanding, proficiency levels of skills, the quality of their products or performances, and their growth from one level to the next" (McTighe & Frontier, 2022, p. 17). A rubric specifies the criteria by which a piece of work will be judged and describes what performance looks like at each level of quality (Andrade, 1997; Brookhart, 2018). When these criteria are closely aligned with the skills being taught and are shared with learners from the start, students can direct their efforts towards the intended outcomes, and the scope for unexamined subjective judgement during grading is narrowed (Andrade & Brookhart, 2026; Jonsson & Svingby, 2007).

Contemporary approaches to teaching, such as independent study, interdisciplinary projects, small-group work, and hands-on inquiry, have made the assessment of learning processes as important as the assessment of recall (Dochy, Gijbels, & Segers, 2006; Styla & Orovou, 2024). Rubrics are frequently presented as well suited to

this type of assessment. They also reflect a wider recognition that a single, uniform approach to testing cannot meet the needs of all students and schools (Benander, Denton, Page, & Skinner, 2000).

Rubrics have already been the subject of several reviews, so a further review requires a clear rationale. Each existing synthesis addresses a narrower question: Jonsson and Svingby (2007) examined reliability and validity; Reddy and Andrade (2010) focused on higher education; Panadero and Jonsson (2013) reviewed formative use; Brookhart and Chen (2015) examined the quality of descriptive rubrics; English, Robertson, Gillis, and Graham (2022) mapped formative use in school education; Panadero, Jonsson, Pinedo, and Fernández-Castilla (2023) meta-analysed effects on performance, self-regulated learning (SRL), and self-efficacy; Jonsson, Panadero, Pinedo, and Fernández-Castilla (2025) analysed implementation factors; and Tian and Tong (2026) reviewed self- and peer-assessment.

However, few reviews have jointly mapped the functions that rubrics serve for students, teachers, and families while weighing the evidence for each function separately. In addition, contradictory findings on reliability, creativity, and the accuracy of self-assessment are usually reported in isolation rather than examined side by side. A scoping design is therefore appropriate, because the aim is to map the extent and nature of a heterogeneous body of literature rather than to estimate a single pooled effect (Arksey & O'Malley, 2005; Tricco et al., 2018). The review addressed two research questions (RQs): RQ1. What is the nature of the evidence on rubric use in education in terms of source types, research designs, and settings? RQ2. What functions do rubrics serve for students, teachers, and families, and how strong is the evidence for each?

Conceptual Background: Rubric Structure and Types

In the Greek-language literature, Kontogiannis (2003) describes rubrics as qualitative performance sheets, while Kouloumbaritsi and Matsagouras (2004) define a rubric as a "graded criteria scale" consisting of four elements (Table 1): (A) evaluation criteria (dimensions), that is, the requirements a task must meet; (B) quality levels (standards), that is, bands of quality ranging from superior to inadequate; (C) a point scale, that is, the numerical value assigned to each level; and (D) level descriptors, that is, written descriptions of what the work must demonstrate at each level for each criterion. A rubric requires at least two criteria and two quality levels.

Table 1. Structure of an analytic rubric (elements A–D)

(A) Evaluation criteria	(B) Level 3 – Superior	(B) Level 2 – Moderate	(B) Level 1 – Low	(C) Points
Criterion 1 (e.g., content accuracy)	(D) Descriptor of superior work on criterion 1	(D) Descriptor of moderate work on criterion 1	(D) Descriptor of low-level work on criterion 1	3 / 2 / 1 (× weight, if weighted)
Criterion 2 (e.g., organisation)	(D) Descriptor of superior work on criterion 2	(D) Descriptor of moderate work on criterion 2	(D) Descriptor of low-level work on criterion 2	3 / 2 / 1 (× weight, if weighted)
Total score				Sum of criterion points

Note. Based on the four-element model of Kouloumbaritsi and Matsagouras (2004). The letters A–D refer to the elements described in the text. Each cell in the (B) columns contains a level descriptor (D), and the final column shows the points (C) awarded for each level.

Rubrics are used to assess interdisciplinary projects, writing, laboratory reports, oral presentations, and group work (Hafner & Hafner, 2003), and they are usually classified as holistic or analytic (Tables 2 and 3). Holistic rubrics evaluate a piece of work as an integrated whole and are used mainly for summative grading. Analytic rubrics divide performance into two or more criteria that are scored separately; they provide more diagnostic feedback and are well suited to multi-stage projects (Allen & Tanner, 2006; Montgomery, 2000). Analytic rubrics can also be weighted to indicate which criteria matter most for a particular task (Taiwo & Vellinga, 2026). Recent evidence suggests that learning outcomes depend less on the format chosen than on how the

criteria are introduced, understood, practised, and applied in self- and peer-assessment (English et al., 2022; Jonsson et al., 2025; Tian & Tong, 2026).

Table 2. Example of a holistic rubric for an oral presentation

Score	Performance description
3 – Superior	Explains the main topic with ease and supports claims with specific detail. Ideas are organised logically. The presenter is clearly well prepared and answers audience questions confidently and clearly.
2 – Proficient	Provides sufficient context and detail, with minor lapses in organisation. Answers most audience questions.
1 – Basic	The main argument lacks strong supporting evidence. Organisation is inconsistent, and the presenter struggles to answer several audience questions.
0 – Unsatisfactory	The topic is unclear, information is disorganised, and the presenter is difficult to follow. No audience questions are answered.

Note. Illustrative example adapted from the Rubistar rubric generator (https://rubistar.4teachers.org).

Table 3. Example of an analytic rubric for a final product presentation

Criterion	4 – Advanced	3 – Proficient	2 – Developing	1 – Unsatisfactory
Preparation	Thoroughly prepared; delivery well rehearsed.	Well prepared; more rehearsal would have helped.	Limited practice; basic preparation evident.	Not prepared for the task.
Content and organisation	Fully on topic; clear introduction, body, and conclusion.	On topic almost throughout; logically structured.	Mostly on topic; sequence sometimes hard to follow.	Focus unclear; audience cannot identify the topic.
Eye contact	Rarely uses notes; steady eye contact throughout.	Checks notes occasionally; frequent eye contact.	Refers to notes constantly; minimal eye contact.	Reads from notes; no eye contact.
Oral expression	Clear, audible, precise vocabulary; no errors.	Clear and audible; minor verbal slips.	Several verbal errors; relies on gestures.	Hard to hear; frequent errors throughout.
Time management	Within the required 5–6 minutes.	About 4 minutes.	About 3 minutes.	Under 3 or over 6 minutes.

Note. Illustrative example adapted from the Rubistar rubric generator (https://rubistar.4teachers.org).

Perlman (2003) identifies two steps in developing a rubric for formative use: designing the task and establishing clear assessment benchmarks. Baartman, Bastiaens, Kirschner, and Van der Vleuten (2006) support a third, post-instructional step, in which the quality of the rubric itself is reviewed. Their "Wheel of Competency Assessment" sets out twelve quality criteria, including reproducibility of decisions, transparency, acceptability, fairness, and authenticity. In a similar vein, Dawson (2017) argues that rubric research and practice are hampered by incomplete reporting of design elements, such as the number of levels, the specificity

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

of the criteria, and the use of exemplars, and he proposes a framework for describing rubrics so that studies can be compared and replicated.

METHOD

Design and Protocol

The review was designed as a systematic scoping review based on the methodological framework of Arksey and O'Malley (2005) and is reported in accordance with the PRISMA extension for scoping reviews (Tricco et al., 2018); the flow diagram follows the PRISMA 2020 template (Page et al., 2021). A scoping design was preferred to a systematic review of effects because the literature is heterogeneous in design and purpose and includes non-empirical sources that nonetheless shape practice. Mapping the evidence and distinguishing between its types is therefore an aim in its own right.

Eligibility Criteria

Eligibility criteria were defined using the Population–Concept–Context (PCC) framework recommended for scoping reviews (Table 4).

Table 4. Eligibility criteria

Dimension	Inclusion	Exclusion
Population / context	Learners, teachers, or families in formal education at any level: K–12, higher education (including teacher education), vocational education and training; face-to-face, blended, or online.	Non-educational settings (e.g., clinical or workplace appraisal) unless the setting is part of a formal education programme.
Concept	Rubrics or scales of graded criteria, i.e., instruments with explicit criteria and descriptors for at least two quality levels, used for assessment, feedback, or instruction.	Checklists or rating scales without quality-level descriptors; publications limited to step-by-step rubric construction without evaluative, conceptual, or outcome content.
Outcomes / focus	Learning, performance, SRL, self-efficacy, self- and peer-assessment, feedback use, scoring reliability or validity, teacher practice, communication with families, or conceptual analysis of rubric functions.	No link to any of the listed outcomes or functions.
Source type	Empirical studies of any design; systematic, scoping, and narrative reviews; conceptual papers; books and chapters; conference papers; theses; institutional reports.	Commercial material, unauthored web pages, and blog posts without institutional or scholarly provenance; editorials without substantive argument.
Language	English or Greek.	All other languages.
Date	January 1997 – July 2026 for database searches. Earlier foundational works identified through citation searching were used as theoretical background only and were not counted in the synthesis corpus.	Pre-1997 sources in the synthesis corpus.
Availability	Full text accessible to the reviewers.	Full text not retrievable.

Information Sources and Search Strategy

Scopus, the Web of Science Core Collection, and ERIC were searched between June and August 2026. Google Scholar served as a supplementary source; because its results cannot be fully exported and are not reproducible, only the first 100 results for each query, sorted by relevance, were screened. Greek-language literature was identified through Google Scholar, Greek institutional repositories, and the proceedings of Panhellenic conferences. Backward citation searching (the reference lists of all included sources) and forward citation searching (citations of the included meta-analyses and reviews) were used to identify material not retrieved by the database searches. The full search strings are reported in Table 5.

Table 5. Search strings by database

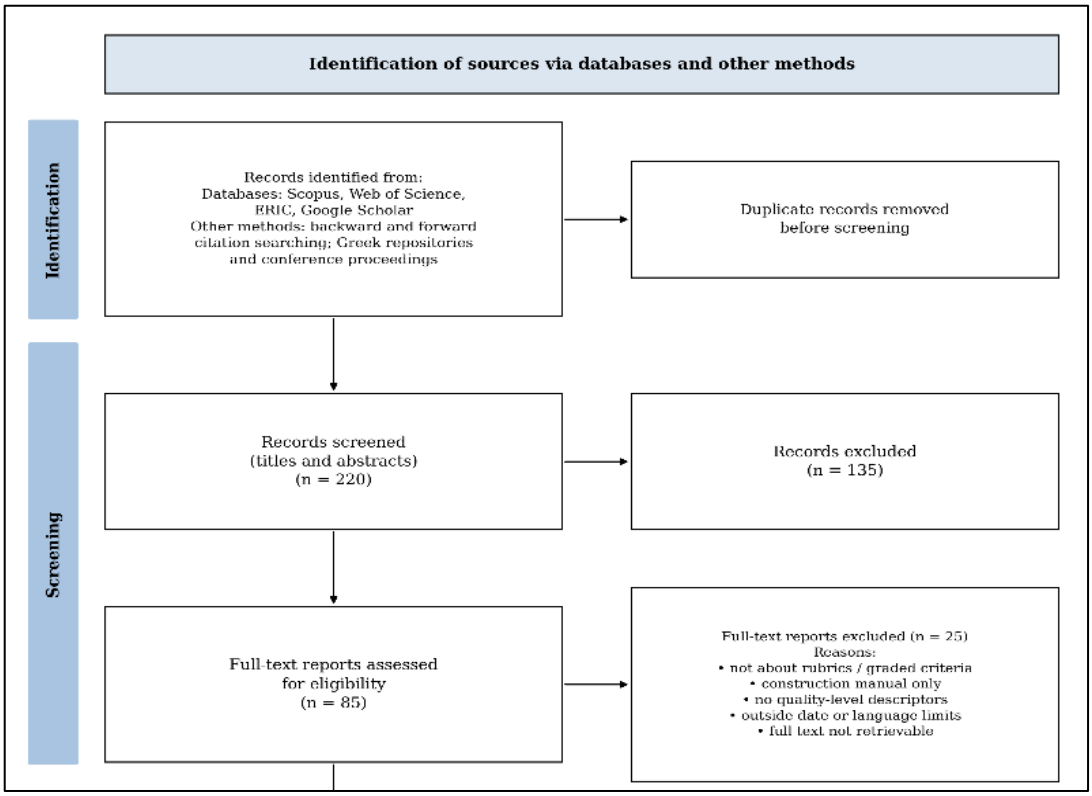
Database	Search string and limits
Scopus	TITLE-ABS-KEY (rubric* OR "scoring guide*" OR "graded criteria" OR "scoring scale*") AND TITLE-ABS-KEY (assess* OR feedback OR "self-assess*" OR "peer assess*" OR "self-regulat*" OR "self-efficacy" OR achievement OR performance OR reliab* OR validity) AND TITLE-ABS-KEY (student* OR learner* OR teacher* OR school* OR classroom* OR universit* OR "higher education" OR vocational) Limits: language = English or Greek; subject area = Social Sciences, Psychology.
Web of Science Core Collection	TS=(rubric* OR "scoring guide*" OR "graded criteria" OR "scoring scale*") AND TS=(assess* OR feedback OR "self-assess*" OR "peer assess*" OR "self-regulat*" OR "self-efficacy" OR achievement OR reliab* OR validity) AND TS=(student* OR teacher* OR school* OR classroom* OR "higher education" OR vocational) Limits: categories = Education & Educational Research, Psychology Educational; language = English.
ERIC	(DE "Scoring Rubrics" OR TI rubric* OR AB rubric*) AND (DE "Student Evaluation" OR DE "Formative Evaluation" OR DE "Self Evaluation (Individuals)" OR DE "Peer Evaluation" OR DE "Interrater Reliability" OR feedback OR "self-regulat*" OR achievement) Limits: journal articles, reports, and conference papers.
Google Scholar (supplementary)	(1) allintitle: rubric OR rubrics assessment; (2) rubric* "self-assessment" OR "self-regulated learning"; (3) rubric* reliability rater training; (4) rubric* "online learning" OR "distance education"; (5) rubric* vocational OR competency. Greek terms: ρουμπρίκες OR "κλίμακες διαβαθμισμένων κριτηρίων" OR "περιγραφική αξιολόγηση" κριτήρια. First 100 results per query screened.

Selection process

Records from all sources were combined, and duplicates were identified and removed by hand. Three reviewers then screened the remaining records manually and independently in two stages, first by title and abstract and then by full text, against the eligibility criteria in Table 4. At each stage, the reviewers compared their decisions, and any disagreements were resolved through discussion until consensus was reached. Reasons for exclusion at the full-text stage were recorded. The selection process is summarised in Figure 1.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

Figure 1. PRISMA 2020 flow diagram



Data Charting and Synthesis

A standardised charting form was piloted on five sources and then applied to all included sources. The form recorded author(s) and year; country; source type; research design; setting (educational level, mode of delivery, discipline); rubric type (holistic/analytic, generic/task-specific, co-constructed or not); the rubric function(s) addressed; outcomes; main findings; and limitations, whether reported by the authors or identified during appraisal. The synthesis combined descriptive mapping with thematic analysis. Themes were first coded deductively from four candidate functions drawn from the literature (communicative, instructional, evaluative, and professional) and were then refined inductively during charting. Within each theme, findings are presented in order of evidence tier. Conflicting findings were tabulated and examined for possible moderating conditions, including design features, implementation conditions, and context.

Evidence Tiers and Quality Appraisal

To distinguish the types of evidence behind each claim, every included source was assigned to one of four evidence tiers according to its design: Tier A, meta-analyses and systematic or scoping reviews; Tier B, experimental and quasi-experimental studies; Tier C, other empirical studies (e.g., correlational, descriptive, qualitative, and action research); and Tier D, narrative reviews, position papers, books, handbooks, and practice reports. The methodological quality of the primary empirical studies (Tiers B and C) was appraised with the Mixed Methods Appraisal Tool (MMAT; Hong et al., 2018). In line with scoping-review methodology, no source was excluded on the grounds of quality; the appraisal was used to qualify the interpretation of findings.

RESULTS

Characteristics of the Evidence Base

Sixty sources were included. Of the 60 sources included in the review, 27 were empirical and 33 were non-empirical, and they were grouped into four evidence tiers by design. Tier A comprised six meta-analyses and systematic or scoping reviews (Brookhart & Chen, 2015; English et al., 2022; Jonsson & Svingby, 2007;

Jonsson et al., 2025; Panadero & Jonsson, 2013; Panadero et al., 2023). Tier B comprised six experimental or quasi-experimental studies (Fraile et al., 2023; Karaman, 2024; Krebs et al., 2022; Rezaei & Lovorn, 2010; Sadler & Good, 2006; Toth et al., 2002). Tier C included 15 other empirical studies (Andrade & Du, 2005; Beeth et al., 2001; Beyreli & Ari, 2009; Bissell & Lemons, 2006; de Barba et al., 2025; Hafner & Hafner, 2003; Hitt & Helms, 2009; Ling, 2025; Luft, 1999; Morrell & Ackley, 1999; Pack et al., 2024; Piscitello, 2001; Rickey et al., 2025; Schamber & Mahoney, 2006; Yavuz et al., 2025). The largest group, Tier D, consisted of 33 narrative reviews, position papers, books, handbooks, and practice reports (Alivizos & Liapis, 2007; Allen & Tanner, 2006; Andrade, 1997, 2005; Andrade & Brookhart, 2026; Arter & McTighe, 2001; Baartman et al., 2006; Baron & Keller, 2003; Benander et al., 2000; Brookhart, 2018, 2024; Chapman & Inman, 2009; Chowdhury, 2019; Dawson, 2017; Gallavan & Kottler, 2009; Hegler, 2003; Jamison, 2000; Kartsiotis & Karatasios, 2008; Kohn, 2006; Kontogiannis, 2003; Kouloumbaritsi & Matsagouras, 2004; Mabry, 1999; McTighe & Frontier, 2022; Montgomery, 2000; Panadero & Jonsson, 2020; Perlman, 2003; Popham, 1997; Reddy & Andrade, 2010; Sadler, 2009; Shaw, 2004; Taiwo & Vellinga, 2026; Tian & Tong, 2026; Volk, 2002). Many of the Tier D sources were published between 1997 and 2009. The most frequent weaknesses of the empirical studies were the lack of a comparison group, reliance on teacher or student self-report, single-site samples, and confounding of the rubric with other interventions, such as instructor feedback.

A Functional Typology of Rubrics

Across the corpus, rubrics were described as serving four distinct but overlapping functions. Table 6 presents these as a conceptual framework, together with the mechanism through which each function is assumed to operate, its main beneficiaries, and the strongest evidence supporting it. The framework highlights a point that is easily lost when benefits are simply listed together: the instructional function is supported by meta-analytic and experimental evidence, the evaluative function by mixed empirical evidence, and the communicative and professional functions mainly by non-empirical (Tier D) sources, with only limited descriptive empirical support. Three groups of moderators cut across all four functions: design features (format, specificity, quality of descriptors, and weighting), implementation conditions (how the rubric is introduced, co-constructed, practised, and linked to feedback), and context (educational level, mode of delivery, discipline, and stakes).

Table 6. Functional typology of rubrics

Function	What the rubric does	Main beneficiaries	Proposed mechanism	Strongest evidence	Key conditions
Communicative	Makes criteria and standards explicit before and during work; creates a shared language	Students, families	Transparency ("feed-up"); reduced uncertainty about expectations	Andrade & Du, 2005	Student-friendly wording; shared at task launch; explained to families
Instructional	Scaffolds self- and peer-assessment, feedback use, goal setting, revision	Students	SRL cycle (planning, monitoring, evaluating); evaluative judgement	Panadero et al., 2023; Krebs et al., 2022; Karaman, 2024	Formative use; explicit training; exemplars; linked to feedback
Evaluative	Standardises judgements for grading and certification	Teachers, institutions	Constrains rater judgement to defined criteria and levels	Jonsson & Svingby, 2007; Rezaei & Lovorn, 2010	Rater training; anchor exemplars; analytic, task-specific format for high stakes
Professional	Supports planning, monitoring progress, and reflection on teaching	Teachers, schools	Criteria act as curriculum-aligned targets and progress data	Hitt & Helms, 2009	Time for design; collegial moderation; curriculum alignment

Instructional Function: Effects on Students

The strongest evidence comes from the meta-analysis by Panadero et al. (2023), which found a moderate positive effect of rubric use on academic performance and smaller positive effects on SRL and self-efficacy. An earlier review concluded that formative use of rubrics may support performance by making expectations transparent, lowering anxiety, strengthening feedback, and fostering self-regulation, although few of the studies reviewed used controlled designs (Panadero & Jonsson, 2013). Experimental evidence indicates that rubrics make students' self-assessments more accurate and lessen the cognitive load involved in self-assessing (Krebs, Rothstein, & Roelle, 2022). In higher education, quasi-experimental studies have reported positive associations between rubric use and SRL, self-efficacy, and performance in collaborative problem-solving (Fraile, Gil-Izquierdo, & Medina-Moral, 2023). Rubric-based self-assessment combined with instructor feedback improved the performance and SRL of pre-service teachers (Karaman, 2024), although the design does not allow the effect of the rubric to be separated from that of the feedback. In secondary science, reflective assessment supported by explicit criteria was associated with a clearer understanding of assessment goals (Toth, Suthers, & Lesgold, 2002).

Qualitative research indicates that students use rubrics to plan, monitor, and revise their work, and that rubrics reduce uncertainty about what teachers expect (Andrade & Du, 2005). Process studies show that students self-assess in diverse ways and do not always refer to the criteria, even when these are available (Rickey, Panadero, & DeLuca, 2025). Teacher reports from primary science and teacher education describe improvements in students' reflection and self-evaluation (Beeth et al., 2001; Luft, 1999), and action research with co-constructed rubrics reports greater commitment to the task (Piscitello, 2001).

Several position papers and practice-oriented texts recommend involving students in generating the criteria, for example through the brainstorming technique described by Osborn (1953) (Andrade, 2005; Andrade & Brookhart, 2026), and Hegler (2003) argues that rubrics develop analytical thinking. Drawing on a narrative review of K–12 studies, Brookhart (2024) concludes that rubrics are most effective when students are explicitly taught how to use them.

Evaluative Function: Reliability and Validity

In their systematic review, Jonsson and Svingby (2007) found that rubrics can enhance scoring reliability, particularly when they are analytic and task-specific and are supplemented by exemplars and rater training. However, they also found that reported reliability frequently fell below the levels required for high-stakes decisions and that validity was seldom investigated. Among primary studies, rubric-guided peer ratings in undergraduate biology showed high consistency (Hafner & Hafner, 2003), and an analytic writing rubric yielded high inter-rater agreement among teachers (Beyreli & Ari, 2009). By contrast, an experimental study found that rubrics did not improve reliability or validity among untrained raters, whose scores continued to be influenced by surface features such as mechanical errors (Rezaei & Lovorn, 2010). Brookhart and Chen (2015) concluded that well-designed descriptive rubrics can support reliable and valid use, but emphasised that the quality of the rubric itself is the limiting factor. Ling (2025) reported adequate reliability for a rubric-based assessment. A more recent line of research concerns rubric-based scoring by large language models (LLMs). Comparative studies report that, in some configurations, LLMs applying a rubric achieve levels of agreement with human raters that approach those observed between human raters; however, agreement varies substantially across models, prompts, and criteria (Pack, Barrett, & Escalante, 2024; Yavuz, Çelik, & Yavaş Çelik, 2025).

Communicative and Professional Functions: Teachers and Families

Claims about the benefits of rubrics for teachers come mainly from handbooks and practice-oriented literature. These sources describe rubrics as helping teachers to define standards, write clearer assignment instructions, grade more consistently, save marking time, and monitor progress (Arter & McTighe, 2001; Chowdhury, 2019; Jamison, 2000; Volk, 2002). Crehan (as cited in Beyreli & Ari, 2009) notes that analytic rubrics give teachers immediate information on whether learning goals are being achieved. The only empirical study in this

area, a descriptive study of in-service teachers, found that teachers perceived rubrics as promoting fairness (Hitt & Helms, 2009).

Communication with families is discussed only in non-empirical sources. Perlman (2003) describes rubrics as a bridge between teachers, students, and parents, and Gallavan and Kottler (2009) and Shaw (2004) argue that sharing rubrics with families reduces suspicions of biased grading and redirects parent–teacher conversations towards strengths and areas for improvement. No included study measured families' understanding, satisfaction, or involvement; these claims therefore remain plausible but unverified. Likewise, the claim that rubrics save teachers time has not been tested. Kontogiannis (2003) and Popham (1997) point to the substantial time required to design and refine good rubrics, which suggests that any time savings may emerge only after an initial investment.

DISCUSSION AND CONCLUSION

This review aimed to separate what is known about rubrics from what is often assumed. Three findings stand out. First, the strength of the evidence differs markedly by function. The instructional function (support for learning, SRL, and self-assessment) is backed by meta-analytic and experimental evidence; the evaluative function by mixed findings that depend on rater training and rubric design; and the communicative and professional functions, including benefits for families and for teacher workload, almost entirely by professional argument. Second, some widely repeated claims are not supported by data. The criticism that rubrics constrain creativity and encourage formulaic work (Chapman & Inman, 2009; Kohn, 2006; Mabry, 1999; Sadler, 2009) has been argued (Panadero & Jonsson, 2020), and rubrics improve the accuracy of students' self-assessment (Krebs et al., 2022) more consistently than they improve the quality of the work itself (Ling, 2025). Third, apparent contradictions in the literature become coherent once moderating conditions are taken into account. Rubrics tend to be beneficial when they are used formatively, when students are taught how to use them, when raters are trained, and when criteria describe qualities of the work rather than prescribe the steps of a task; they are much less helpful when they are simply handed out as scoring sheets. This supports, and makes more specific, the conclusion of Jonsson et al. (2025) that implementation, rather than the rubric itself, determines outcomes. The typology in Table 6 provides a way of organising future research and practice around these distinctions. Rather than asking "Do rubrics work?", it is more useful to ask which function a given rubric is intended to serve, for whom, and under what conditions. A rubric designed mainly for reliable certification (analytic, task-specific, and used by trained raters) may differ considerably from one designed mainly to develop students' evaluative judgement (generic, co-constructed, and used repeatedly across tasks).

Implications for rater training: Because rubrics do not by themselves ensure reliable scoring (Rezaei & Lovorn, 2010), and because rater training improves consistency (Jonsson & Svingby, 2007; Weigle, 1998), schools and departments that use rubrics for grades that count should (a) hold a calibration session before each marking period, in which raters score a set of anchor samples (at least one per quality level) and discuss discrepancies until they agree on how the descriptors should be applied; (b) before independent marking begins, double-mark a sample of work (e.g., 10–20%) and calculate an agreement index; (c) explicitly train raters to disregard construct-irrelevant features, such as handwriting, length, or mechanical errors, when these are not among the criteria.

Implications for student training: Because the learning benefits of rubrics depend on students' active use of them (Brookhart, 2024; Panadero et al., 2023), and because understanding the criteria does not automatically lead to better work (Ling, 2025), teachers should (a) introduce each rubric together with annotated exemplars at different quality levels, rather than presenting the rubric on its own; (b) ask students to score the exemplars and compare their judgements with the teacher's before they assess their own work; (c) involve students in rewording or co-constructing the descriptors in their own words, which fosters ownership and evaluative judgement (Andrade & Brookhart, 2026; Tai et al., 2018); (d) require a rubric-based self-assessment before submission and give feedback on the accuracy of the self-assessment itself, not only on the work (Karaman, 2024; Krebs et al., 2022); and (e) adapt rubrics to students' age, using two or three criteria.

Implications for curriculum planning and families: At school level, rubrics can be aligned with curriculum outcomes and organised into a shared bank in which criteria progress coherently across grades (Arter & McTighe, 2001), supported by regular moderation meetings among teachers. When rubrics are shared with families, a brief explanation of the criteria and quality levels is likely to be needed. Schools that do so could collect simple data on parents' understanding, which would begin to close the current evidence gap.

Generally, the functional typology proposed here, together with the moderators it identifies, gives teachers a basis for deciding what a rubric is for and how to implement it, and gives researchers a map of where evidence is still needed.

Limitations and Directions For Future Research

Several limitations should be acknowledged. The search was restricted to English- and Greek-language sources, so relevant work in other languages, particularly Spanish-language research on rubrics, may have been missed. Google Scholar searches are not fully reproducible, although Google Scholar was used only as a supplementary source. The evidence tiers classify sources by design, and a design-based classification cannot capture every aspect of study quality. Finally, as a scoping review, this study did not pool effect sizes. Future research should move beyond asking whether rubrics "work" and should examine under which conditions, for whom, and through which mechanisms they are effective.

Ethical Considerations

As this study is a review of published literature and involved no human participants, animal subjects, or collection of primary data, ethical approval was not required.

Conflict Of Interest

The author(s) declare no conflict of interest.

Data Availability

No new data were generated or analysed in this study.

REFERENCES

1. Alivizos, S., & Liapis, V. (2007). Innovations in contemporary educational reality: Assessment using scales of graded criteria (rubrics) and student self-assessment using learning logs in media education. In Proceedings of the 4th Panhellenic Conference of EEEP–DTPE "ICT & Education". [In Greek]
2. Allen, D., & Tanner, K. (2006). Rubrics: Tools for making learning goals and evaluation criteria explicit for both teachers and learners. CBE—Life Sciences Education, 5(3), 197–203. https://doi.org/10.1187/cbe.06-06-0168
3. Andrade, H. G. (1997). Understanding rubrics. Educational Leadership, 54(4), 14–17.
4. Andrade, H. G. (2005). Teaching with rubrics: The good, the bad, and the ugly. College Teaching, 53(1), 27–31. https://doi.org/10.3200/CTCH.53.1.27-31
5. Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. Practical Assessment, Research, and Evaluation, 10(3).
6. Andrade, H. L., & Brookhart, S. M. (2026). Using rubrics for teaching and learning. Routledge.
7. Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. International Journal of Social Research Methodology, 8(1), 19–32. https://doi.org/10.1080/1364557032000119616
8. Arter, J., & McTighe, J. (2001). Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance. Corwin Press.
9. Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. Studies in Educational Evaluation, 32(2), 153–170. https://doi.org/10.1016/j.stueduc.2006.04.006

10. Baron, J., & Keller, M. (2003, November). Use of rubrics in online assessment [Paper presentation]. Evaluations and Assessment Conference, University of South Australia, Adelaide, Australia.
11. Beeth, M. E., Cross, L., Pearl, C., Pirro, J., Yagnesak, K., & Kennedy, J. (2001). A continuum for assessing science process knowledge in grades K–6. Electronic Journal of Science Education, 5(3).
12. Benander, R., Denton, J., Page, D., & Skinner, C. (2000). Primary trait analysis: Anchoring assessment in the classroom. The Journal of General Education, 49(4), 279–302.
13. Beyreli, L., & Ari, G. (2009). The use of analytic rubric in the assessment of writing performance: Inter-rater concordance study. Educational Sciences: Theory and Practice, 9(1), 105–125.
14. Bissell, A. N., & Lemons, P. P. (2006). A new method for assessing critical thinking in the classroom. BioScience, 56(1), 66–72.
15. Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. Frontiers in Education, 3, Article 22. https://doi.org/10.3389/feduc.2018.00022
16. Brookhart, S. M. (2024). Using rubrics in basic education: A review and recommendations. Estudos em Avaliação Educacional, 35, Article e10803. https://doi.org/10.18222/eae.v35.10803
17. 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
18. Chapman, V. G., & Inman, M. D. (2009). A conundrum: Rubrics or creativity/metacognitive development? Educational Horizons, 87(3), 198–202.
19. Chowdhury, F. (2019). Application of rubrics in the classroom: A vital tool for improvement in assessment, feedback and learning. International Education Studies, 12(1), 61–68. https://doi.org/10.5539/ies.v12n1p61
20. Dawson, P. (2017). Assessment rubrics: Towards clearer and more replicable design, research and practice. Assessment & Evaluation in Higher Education, 42(3), 347–360. https://doi.org/10.1080/02602938.2015.1111294
21. de Barba, P. G., Oliveira, E. A., & English, N. (2025). Development and validation of a learning analytics rubric for self-regulated learning. Educational Technology Research and Development, 73, 3223–3245. https://doi.org/10.1007/s11423-025-10521-x
22. Dochy, F., Gijbels, D., & Segers, M. (2006). Learning and the emerging new assessment culture. In L. Verschaffel, F. Dochy, M. Boekaerts, & S. Vosniadou (Eds.), Instructional psychology: Past, present and future trends (pp. 191–206). Elsevier.
23. English, N., Robertson, P., Gillis, S., & Graham, L. (2022). Rubrics and formative assessment in K–12 education: A scoping review of literature. International Journal of Educational Research, 113, Article 101964. https://doi.org/10.1016/j.ijer.2022.101964
24. Fraile, J., Gil-Izquierdo, M., & Medina-Moral, E. (2023). The impact of rubrics and scripts on self-regulation, self-efficacy and performance in collaborative problem-solving tasks. Assessment & Evaluation in Higher Education, 48(8), 1223–1239. https://doi.org/10.1080/02602938.2023.2236335
25. Gallavan, N. P., & Kottler, E. (2009). Constructing rubrics and assessing progress collaboratively with social studies students. The Social Studies, 100(4), 154–158.
26. Hafner, J. C., & Hafner, P. M. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. International Journal of Science Education, 25(12), 1509–1528.
27. Hegler, K. L. (2003, January 24–27). Using general education assessment rubrics to document basic skills and content knowledge [Paper presentation]. 55th Annual Meeting of the American Association of Colleges for Teacher Education, New Orleans, LA, United States. ERIC Document Reproduction Service No. ED472812.
28. Hitt, A. M., & Helms, E. C. (2009). Best in show: Teaching old dogs to use new rubrics. The Professional Educator, 33(1), 1–15.
29. Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Education for Information, 34(4), 285–291. https://doi.org/10.3233/EFI-180221
30. Jamison, P. (2000). Student evaluation: A staff development handbook. Saskatchewan Education and Saskatchewan Professional Development Unit.

31. Jonsson, A., Panadero, E., Pinedo, L., & Fernández-Castilla, B. (2025). Using rubrics for formative purposes: Identifying factors that may affect the success of rubric implementations. Assessment in Education: Principles, Policy & Practice, 32(2), 192–211. https://doi.org/10.1080/0969594X.2025.2486947
32. Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review, 2(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002
33. Karaman, P. (2024). Effects of using rubrics in self-assessment with instructor feedback on pre-service teachers' academic performance, self-regulated learning and perceptions of self-assessment. European Journal of Psychology of Education, 39, 2551–2574. https://doi.org/10.1007/s10212-024-00867-w
34. Kartsiotis, T., & Karatasios, G. (2008). The pedagogical utilization of ICT in education. In Proceedings of the 1st Panhellenic Conference of Imathia "Digital material for supporting the pedagogical work of teachers" (pp. 16–25). [In Greek]
35. Kohn, A. (2006). The trouble with rubrics. English Journal, 95(4), 12–15.
36. Kontogiannis, K. (2003). The basic characteristics of qualitative descriptive assessment. e-paideia.net. [In Greek]
37. Kouloumbaritsi, A. C., & Matsagouras, I. G. (2004). The student's portfolio (portfolio assessment): Authentic assessment in theory and practice. Pedagogical Review, 40. [In Greek]
38. Krebs, R., Rothstein, B., & Roelle, J. (2022). Rubrics enhance accuracy and reduce cognitive load in self-assessment. Metacognition and Learning, 17(2), 627–650. https://doi.org/10.1007/s11409-022-09302-1
39. Ling, J. H. (2025). Assessing the validity and reliability of a rubric-based assessment. Journal of Multidisciplinary Research and Development, 2(2). https://ejournal.papanda.org/index.php/jmrd/article/view/1253
40. Luft, J. A. (1999). Rubrics: Design and use in science teacher education. Journal of Science Teacher Education, 10(2), 107–121. https://doi.org/10.1023/A:1009471931127
41. Mabry, L. (1999). Writing to the rubric: Lingering effects of traditional standardized testing on direct writing assessment. Phi Delta Kappan, 80(9), 673–679.
42. McTighe, J., & Frontier, T. (2022). How to provide better feedback through rubrics. https://jaymctighe.com/wp-content/uploads/2022/04/How-To-Provide-Better-Feedback-Through-Rubrics.pdf
43. Montgomery, K. (2000). Classroom rubrics: Systematizing what teachers do naturally. The Clearing House, 73(6), 324–328.
44. Morrell, P. D., & Ackley, B. C. (1999, April). Practicing what we teach: Assessing pre-service teachers' performance using scoring guides [Paper presentation]. Annual Meeting of the American Educational Research Association, Montreal, Canada.
45. Osborn, A. F. (1953). Applied imagination: Principles and procedures of creative problem-solving. Scribner's.
46. Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. Computers and Education: Artificial Intelligence, 6, Article 100234. https://doi.org/10.1016/j.caeai.2024.100234
47. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71
48. Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. Educational Research Review, 9, 129–144. https://doi.org/10.1016/j.edurev.2013.01.002
49. 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
50. 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, Article 113. https://doi.org/10.1007/s10648-023-09823-4

51. Perlman, C. C. (2003). Performance assessment: Designing appropriate performance tasks and scoring rubrics. In J. E. Wall & G. R. Walz (Eds.), Measuring up: Assessment issues for teachers, counselors, and administrators (pp. 497–506). ERIC Counseling and Student Services Clearinghouse. ERIC Document Reproduction Service No. ED480070.
52. Piscitello, M. E. (2001). Using rubrics for assessment and evaluation in art [Master's thesis, Saint Xavier University].
53. Popham, W. J. (1997). What's wrong—and what's right—with rubrics. Educational Leadership, 55(2), 72–75.
54. Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. Assessment & Evaluation in Higher Education, 35(4), 435–448. https://doi.org/10.1080/02602930902862859
55. Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. Assessing Writing, 15(1), 18–39. https://doi.org/10.1016/j.asw.2010.01.003
56. Rickey, N., Panadero, E., & DeLuca, C. (2025). How do students self-assess? Examining the metacognitive processes of student self-assessment. Metacognition and Learning, 20(1). https://doi.org/10.1007/s11409-025-09430-4
57. Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. Assessment & Evaluation in Higher Education, 34(2), 159–179. https://doi.org/10.1080/02602930801956059
58. Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. Educational Assessment, 11(1), 1–31. https://doi.org/10.1207/s15326977ea1101_1
59. Schamber, J. F., & Mahoney, S. L. (2006). Assessing and improving the quality of group critical thinking exhibited in the final projects of collaborative learning groups. The Journal of General Education, 55(2), 103–137.
60. Shaw, J. (2004). Demystifying the evaluation process for parents: Rubrics for marking student research projects. Teacher Librarian, 32(2), 16–19.
61. Styla, D., & Orovou, E. (2024). Contemporary teaching methodology in education: Theory and practice. Konstantaras Publications. [In Greek]
62. 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
63. Taiwo, F. A., & Vellinga, E. (2026). Assessment rubrics: Design, development and implementation. In G. Saunders-Smits, L. Van den Broeck, & T. Johannsen (Eds.), SEFI handbook on teaching transferable competencies and skills in engineering. TU Delft OPEN Publishing.
64. Tian, W., & Tong, Y. (2026). A literature review of effects of rubric on student self-assessment and peer assessment in higher education. Journal of Language, Culture and Education, 3(2), 1–13.
65. Toth, E. E., Suthers, D. D., & Lesgold, A. M. (2002). "Mapping to know": The effects of representational guidance and reflective assessment on scientific inquiry. Science Education, 86(2), 264–286. https://doi.org/10.1002/sce.10004
66. Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Annals of Internal Medicine, 169(7), 467–473. https://doi.org/10.7326/M18-0850
67. Volk, J. (2002). Assessment strategies. Saskatchewan Learning–Region 3.
68. Weigle, S. C. (1998). Using FACETS to model rater training effects. Language Testing, 15(2), 263–287. https://doi.org/10.1177/026553229801500205
69. Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. British Journal of Educational Technology, 56(1), 150–166. https://doi.org/10.1111/bjet.13494